UQAC
Université du Québec
à Chicoutimi

EFFECTIVE FAKE NEWS MITIGATION: INTEGRATING XAI, LLMS, AND COGNITIVE EYE-TRACKING ANALYSIS

PAR RACHEL LADOUCEUR

THESIS PROPOSAL PRESENTED TO L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI IN PARTIAL FULFILLMENT OF THE REQUIERMENTS FOR THE DEGREE OF PHILOSOPHIÆ DOCTOR (PH.D.) IN THE SUBJECT OF INFORMATIQUE

QUÉBEC, CANADA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AI	Artificial Intelligence
ALBERT	A lite BERT
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
COMET	Crosslingual Optimized Metric for Evaluation of Translation
CQEMI	Quebec Center for Media and Information Education
DDoS	Denial-of -Service Distributed
DSA	Digital Services Act
EDMO	European Digital Media Observatory
EEAS	European External Action Service
EEG	Electroencephalography
ENISA	European Union Agency for Cybersecurity
FIMI	Foreign Information Manipulation and Interference
GAN	Generative Adversarial Network
IRA	Internet Research Agency
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
LSTM	Long Short-Term Memory
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MLM	masked language modeling
NLP	Natural Language Processing
PPE	Political Person Entities
RNN	Recurrent Neural Network
RoBERTa	Robustly Optimized BERT Pretraining Approach
SHAP	Shapley Additive Explanations
TF-IDF	Term Frequency-Inverse Document Frequency
WHO	World Health Organization
XAI	EXplainable Artificial Intelligence

ACKNOWLEDGEMENTS

I would first like to express my sincere gratitude to my thesis supervisor, **Professor Fehmi Jaafar**, for his rigorous guidance, availability, and invaluable advice throughout this doctoral research. His scientific expertise and continuous support were essential to the successful completion of this work.

I would also like to thank **Mouvement Desjardins** for their financial support, particularly for covering the expenses related to the publication of the scientific articles resulting from this thesis.

I am also grateful to the collaborators who contributed to the various articles associated with this work.

For the first contribution, entitled *Can Deep Learning Detect Fake News Better When Adding Context Features?*, I would like to thank Abubakr Hassan and Professor Mohamed Mejri (Department of Computer Science and Software Engineering, Université Laval) for their contributions.

For the second contribution, *Comparisons Between Open-LLMs for Fake News Detection*, I would like to thank Chaima Njeh and Associate Professor Haïfa Nakouri (University of Tunis – Higher Institute of Management (ISG), LARODEC, Tunisia) for their contributions.

For the third contribution, *Elements of Credibility or Credulity About Fake News Exposure: An Empirical Study*, I would like to thank Tom Pierre and Professor Hamdi Ben Abdessalem (Department of Computer Science and Mathematics, UQAC) for their contributions.

Finally, I would like to thank everyone who indirectly contributed to the completion of this thesis.

OTHER PUBLICATIONS RELATED TO THIS THESIS

BOOK CHAPTERS

Ladouceur, R. (2023). *Cyber Influence Stakes*. In F. Jaafar & S. Pierre (Eds.), *Blockchain and Artificial Intelligence-Based Solutions to Enhance Privacy in Digital Identity and IoT* (pp. 155–174). CRC Press [3].

Ladouceur, R., Pierre, S., & Jaafar, F. (2022). *Cyberwarfare: Geopolitics of Technological Infrastructures and Social Networks*. In D. Uzunidis & L. Adatto (Eds.), *Major Catastrophes in the 21st Century: Health, Environment, Food, War* (Collection Magna Carta). Editions Le Manuscrit [4].

The author would like to express sincere gratitude to **Professor Fehmi Jaafar** for his guidance and support, as well as to **Schallum Pierre, Ph.D.** for his valuable contributions and collaboration throughout this research work.

CONFERENCE CONTRIBUTIONS

Artificial Intelligence and Cyber Influence, ACFAS Congress (Annual Congress of the Francophone Association for the Advancement of Knowledge), May 2022.

ABSTRACT

This thesis investigates the cognitive and computational mechanisms that shape susceptibility to fake news by integrating machine learning, large language models, and eye-tracking analysis.

The first contribution evaluates how contextual metadata can enhance BERT and LSTM models for fake news detection. The findings reveals that enriching a BERT-based architecture with contextual metadata improves the automated fake news detection. It also introduces SHAP which quantifies the contribution of each metadata attribute to the final prediction. SHAP analysis revealed that several metadata features played a significant role in the model's predictions.

The second contribution systematically evaluates the boundaries of large language models to detect and justify fake news within a restrictive zero-shot framework. The findings reveal that when relying solely on internal pretrained knowledge, current LLMs achieve modest detection performance, marginally above chance. Crucially, multi-metric evaluations using BERTScoreF-1 and COMET expose a significant semantic gap: while LLM-generated justifications exhibit surface-level textual coherence, they lack the factual verification, source-based evidence, and semantic adequacy required for reliable characterization. This empirical refutation highlights the critical risks of black-box generative explanations.

The third contribution examines the cognitive dimension of fake news through an eye-tracking experiment. The findings reveal that both the topic of a news item and the presence of Political Person Entities (PPE) significantly shape readers' credulity and visual attention. PPE act as powerful heuristic cues that reduce analytical scrutiny, showing how linguistic features embedded in fake news can exploit cognitive shortcuts. These observations provide empirical insight into the psychological mechanisms underlying vulnerability to fake news and open the way to mitigation strategies that are more sensitive to cognitive processes.

Together, these contributions support the central paradigm of this thesis: effective fake news mitigation requires the joint optimization of computational performance and explicit explainability while structurally accounting for human cognitive limitations. By unifying machine learning interpretability and neurophysiological insights, this dissertation establishes a comprehensive framework for advancing cognitive security in digital information ecosystems.

RÉSUMÉ

Cette thèse examine les mécanismes cognitifs et computationnels qui influencent la vulnérabilité des individus face aux fausses nouvelles, en intégrant des approches issues de l'apprentissage automatique, des grands modèles de langage et de l'oculométrie.

La première contribution évalue comment les métadonnées contextuelles peuvent améliorer les modèles BERT et LSTM pour la détection de fausses informations. Les résultats montrent qu'enrichir une architecture basée sur BERT avec des métadonnées contextuelles améliore la détection automatisée des fausses informations. Elle introduit également SHAP, qui quantifie la contribution de chaque attribut de métadonnée à la prédiction finale. L'analyse SHAP a révélé que plusieurs caractéristiques de métadonnées ont joué un rôle important dans les prédictions du modèle.

La deuxième contribution évalue de manière systématique les limites des grands modèles de langage pour détecter et justifier les fausses nouvelles dans un cadre restrictif de zero-shot. Les résultats montrent que lorsqu'ils s'appuient uniquement sur leurs connaissances préentraînées internes, les modèles actuels obtiennent des performances de détection modestes, à peine supérieures au hasard. Fait crucial, des évaluations multi-métriques utilisant BERTScore-F1 et COMET mettent en évidence un important fossé sémantique : bien que les justifications générées par les LLM présentent une cohérence textuelle, elles manquent de vérification factuelle, de preuves fondées sur des sources et d'adéquation sémantique nécessaires à une caractérisation fiable. Cette réfutation empirique souligne les risques critiques associés aux explications génératives opaques.

La troisième contribution explore la dimension cognitive de la désinformation à l'aide d'une étude oculométrique. Les résultats révèlent que le sujet d'une nouvelle et la présence de noms de figures politiques (PPE) influencent fortement la crédulité des lecteurs et leurs schémas attentionnels. Les PPE agissent comme des indices heuristiques puissants, réduisant l'analyse approfondie et facilitant l'acceptation intuitive de contenus trompeurs. Ces observations apportent un éclairage empirique sur les mécanismes psychologiques qui sous-tendent la vulnérabilité à la désinformation et ouvrent la voie à des stratégies de mitigation plus sensibles aux processus cognitifs.

Ensemble, ces contributions soutiennent le paradigme central de cette thèse : une atténuation efficace des fausses nouvelles exige l'optimisation conjointe des performances computationnelles et de l'explicabilité explicite, tout en tenant structurellement compte des limites cognitives humaines. En unifiant l'interprétabilité en apprentissage automatique et les perspectives neurophysiologiques, cette dissertation établit un cadre global pour faire progresser la sécurité cognitive au sein des écosystèmes numériques de l'information.

CHAPTER I

INTRODUCTION

This chapter introduces the growing complexity of disinformation, tracing its evolution from the popularization of the term fake news during the 2016 U.S. presidential election to the broader concept of Foreign Information Manipulation and Interference (FIMI) adopted by ENISA. It highlights the societal, democratic, and public-health risks amplified by social media ecosystems and recent platform policy changes. The chapter then identifies key research gaps related to human cognitive biases, the limitations of current explainability methods, and the emerging but insufficiently understood capabilities of large language models (LLMs). These gaps motivate the dissertation's three contributions. The chapter concludes by presenting the overarching hypothesis that combining Explainable AI (XAI), LLMs, and cognitive analysis, and outlines the structure of the dissertation.

1.1. THE CONTEXT

The term **"fake news"** gained widespread public attention during the 2016 U.S. presidential campaign, notably through its repeated use by Donald Trump. In its 2025 work, the European Union Agency for Cybersecurity (ENISA), in coordination with the European External Action Service (EEAS), adopts the broader concept of Foreign Information Manipulation and Interference (FIMI) instead of focusing solely on misinformation/disinformation as it was earlier. The new framework highlights the potential impact of such activities on democratic processes, recognizing that not all false or misleading information is harmful, whereas coordinated manipulative behaviour can undermine institutions even when the content itself is not illegal [5].

Throughout this dissertation, we use the term fake news. While the expression has been criticized for its ambiguity and politicization in policy-oriented and institutional contexts, it remains widely used in the academic literature on information diffusion and automated detection. Seminal works such as Vosoughi et al. [6] and subsequent research in machine learning and natural language processing (e.g., Shu et al. [7]) consistently employ the term fake news to denote verifiable false or misleading news. In order to remain consistent with this body of literature, as well as with the terminology used in the datasets and models analyzed in this dissertation, we adopt the term fake news as an operational label, without ignoring its conceptual limitations.

Fake news can erode trust in democratic institutions and serve military and geopolitical objectives, but also be life-threatening, as evidenced during the COVID-19 pandemic. The World Health Organization (WHO) identified vaccine hesitancy, partly fueled by conspiracy theories, as a major threat to global public health [8].

While many instances of fake news remain largely unnoticed, others attract disproportionate attention, generate sustained engagement, and spread rapidly through online platforms. It is these highly amplified messages, often embedded within recurring narrative patterns, that tend to produce the most severe negative impacts on public trust, democratic processes, and public health. The COVID-19 pandemic provides a particularly salient illustration, where a limited number of highly engaging narratives surrounding vaccination had far-reaching consequences.

Multiple studies have shown that social networks are highly effective at rapidly disseminating fake news to large audiences, thereby facilitating the manipulation of opinions and behaviours [9]. One key factor contributing to this effectiveness is the low cost and scalability of online influence operations. This phenomenon has intensified over time: the number of

states engaging in social media fake news campaigns increased from 28 in 2017 to 70 in 2019, nearly quadrupling in just two years [10].

Several measures are being implemented to combat the spread of fake news including media literacy initiatives and platform-based interventions. Recent developments have further complicated the landscape. In January 2025, Meta announced the end of its third-party fact-checking program in the United States, opting for a community-based moderation model similar to that of platform X [11]. This decision drew substantial criticism, as some observers argued that it could facilitate the spread of fake news across its platforms. Faced with this threat, legislators attempt to build defenses, but regulation struggles to keep pace with technological evolution.

The World Economic Forum's 2026 Global Risks Report ranks **disinformation as the second most significant global risk over the next two years**, highlighting its continued severity. This risk is driven primarily by **ongoing armed conflicts** and the **rapid advancement of generative artificial intelligence**, which enables the large-scale creation and dissemination of misleading or manipulated content.

In 2022, ENISA reported disinformation and misinformation as one of the top ten global cyber threats [12]. Since then, the threat has evolved and become more complex. In its 2025 edition, they explicitly highlighting how generative AI multiplies its reach and sophistication. **Information manipulation now appears among the eight top threats, alongside ransomware and Distributed Denial-of-Service (DDoS) attacks** [5].

1.2. RESEARCH GAPS AND MOTIVATIONS

This section presents the main research gaps and motivations associated with the three contributions of this dissertation. The following subsections detail, for each contribution, the identified challenges and the underlying objectives that guide them.

1.2.1. DEEP NEURAL NETWORK AND XAI

Motivation 1. Contextual metadata (e.g., number of followers, hashtags, number of likes, etc.) are features to be captured to detect fake news [7]. Studies used at least two deep neural network to perform fake news detection when jointly analyzed content-based features at the same time [13], which increase model complexity. Moreover, users require meaningful explanations to better understand and critically assess automated decisions. However, existing explanation techniques remain limited. In particular, keyword-based explanations are often weakly aligned with the underlying decision mechanisms of the model and fail to capture the semantic and contextual factors driving fake news detection. This limitation highlights the need for more expressive and feature-aware Explainable AI (XAI) approaches.

Identified Research Gaps. The first contribution proposes a unified (Bidirectional Encoder Representations from Transformers) BERT-based framework that integrates both textual and contextual features to detect fake news. In addition, SHapley Additive exPlanations (SHAP) are employed to quantify the contribution of the metadata to the detection outcome. Compared to Local Interpretable Model-agnostic Explanations (LIME)-based keyword extraction, this approach provides richer, more semantically meaningful explanations that are better aligned with the model's internal decision process [14].

1.2.2. LARGE LANGUAGE MODELS FOR FAKE NEWS DETECTION

Motivation 2. Recent studies have shown that Large Language Models (LLMs) can perform fake news detection using prompting techniques, without requiring task-specific fine-tuning [15, 16, 17, 18]. This capability is enabled by large-scale pretraining on diverse corpora, allowing LLMs to acquire broad world knowledge and general reasoning abilities [19]. However, it remains unclear how the performance of LLMs compares to that of established encoder-based models such as BERT, particularly when both predictive accuracy and explainability are considered in the fake news domain.

Identified Research Gaps. The second contribution evaluates LLMs on fake news detection and assesses the quality of their generated explanations. It presents a comprehensive evaluation of three open-source LLMs in zero-shot settings, for detection performance and explanation quality. In contrast to prior work, explanation quality is quantitatively evaluated using transformer-based semantic similarity metrics such as BERTScore-F1 and COMET, representing a significant methodological advancement over traditional metrics such as METEOR.

1.2.3. EYE-TRACKING ANALYSIS FOR FAKE NEWS

Motivation 3. Human cognition remains central to the information evaluation process, yet individuals are particularly vulnerable to misleading information [20]. This observation aligns with the concept of truth bias, whereby individuals tend to accept information as true unless compelling contradictory evidence is provided [21]. As a result, the skeptical evaluation of news content does not occur naturally and often requires deliberate cognitive effort.

Eye-tracking research has consistently shown that the cognitive processing of fake news differs from that of real news, with misleading information eliciting longer reading times, increased fixation counts, and distinct gaze patterns [22, 23, 24]. Previous studies have also shown that specific words or linguistic cues can capture and sustain visual attention in different ways [25, 26]. However, existing work has primarily focused on global eye-movement metrics and has not systematically examined which specific semantic elements within a claim trigger these cognitive responses [27].

Identified Research Gaps. This third contribution investigates whether attention biases vary depending on the topic and the presence of Political Proper Entities (PPE), or whether they reflect a more general cognitive vulnerability. This contribution provides new insights into cognitive and attentional mechanisms that can support the development of explainability and information literacy tools. This is the first study that explored the presence of PPE such as Obama, Trump, Musk and Biden.

1.3. DISSERTATION HYPOTHESIS

General Hypothesis. This dissertation hypothesizes that AI-based and cognitive approaches including explainable artificial intelligence (XAI), large language models (LLMs), and cognitive eye-tracking analysis improve fake news mitigation.

To answer this hypothesis, three sub-hypotheses are formulated:

Sub-Hypothesis 1. Augmenting a BERT-based architecture with contextual metadata (e.g., number of followers, hashtags, engagement metrics), combined with SHAP-based feature attribution, improves both predictive fake news performance and explainability.

We explored whether metadata provide external credibility signals that are not captured by textual content alone. In this contribution, a large-scale COVID-19 dataset is partially used to evaluate this assumption. In addition, SHAP techniques are used to estimate the relative contribution of contextual attributes, thereby improving the transparency and explainability of the model's predictions.

Sub-Hypothesis 2. Zero-shot prompted large language models can effectively detect fake news and generate explanations that contribute to fake news mitigation.

We explored the boundaries of these models' explanatory and analytical capabilities using the specialized Fin-Fact dataset. This contribution aims to evaluate whether LLMs, acting as knowledge-rich transformers, can independently decode complex financial claims using only their internal pretrained representations, specifically in the absence of explicit social metadata. By measuring the lexical and semantic convergence of automated natural language justifications against human-annotated benchmarks, this contribution formalizes and characterizes the semantic gap between automated heuristic reasoning and human verification rationale. Ultimately, testing this hypothesis allows us to map the operational limitations of standalone generative models and empirically justify the necessity of complementary, transparent XAI frameworks.

Sub-Hypothesis 3. Cognitive eye-tracking analysis explains how people are influenced by fake news.

This contribution examines whether Political Person Entities (PPE) such as Trump, Biden, and Obama, attract disproportionate visual attention and whether this affects cognitive processing and increases the likelihood of misjudgment. Eye-tracking analysis is therefore used to examine whether attention patterns are systematically drawn toward these PPE and whether this effect reflects a broader cognitive vulnerability in fake news evaluation.

1.4. STRUCTURE OF THE DISSERTATION

The remainder of this dissertation is organized as follows.

Chapter 2 surveys deep learning-based approaches for fake news detection, including recurrent neural networks (RNNs) and long short-term memory networks (LSTMs). It then examines the emergence of Transformer-based models, particularly BERT and its variants, which have become the dominant paradigm for content-based fake news detection due to their ability to capture rich contextual and semantic representations. The chapter further reviews explainable artificial intelligence (XAI) techniques developed to improve the transparency, explainability, and trustworthiness of Transformer-based models. Recent advances involving large language models (LLMs) are also examined, with particular emphasis on their potential for both fake news detection and explanation generation. Finally, the chapter discusses cognitive eye-tracking studies used to analyze attention, cognitive load, and emotional engagement during fake news processing.

Chapter 3 presents a comprehensive review of related work on fake news detection, XAI, and cognitive eye-tracking analysis. It also outlines the different algorithms employed in prior studies, ranging from RNNs and BERT-based architectures to LLMs. This review aims to identify methodological trends, limitations, and open research gaps that motivate the contributions of this dissertation.

Chapters 4, 5, and 6 present the methodologies and experimental results corresponding to the three main contributions that constitute the core of this dissertation.

Finally, the conclusion revisits the dissertation hypothesis and the sub-hypotheses associated with the three main contributions. It also synthesizes the limitations identified across the proposed approaches and outlines directions for future research.

Together, these perspectives establish the multidisciplinary foundation of this research and highlight the research gaps motivating the integration of XAI, LLMs, and cognitive eye-tracking analysis for explainable fake news mitigation.

CHAPTER II

BACKGROUND

As early as the 5th century before the Common Era, Sun Tzu emphasized that deception and the manipulation of information are fundamental strategic weapons. In The Art of War, he argued that misleading an opponent is essential to gaining efficient control and securing victory [28].

With the invention of the printing press, manipulation of information and the dissemination of misleading or false claims began to spread on a much larger scale. It has historically flourished particularly during periods of war or political crisis. By the end of the nineteenth century, it had become industrialized. In 1898, American newspapers attributed the sinking of the United States Ship Maine to Spain without conclusive evidence, a narrative that directly contributed to the outbreak of the Spanish–American War. This episode is widely regarded as one of the earliest documented cases in which false or misleading news decisively influenced the initiation of an armed conflict [29, 30].

The twentieth century marked the rise of mass propaganda through radio and television. During the Cold War, major powers conducted large-scale disinformation campaigns. One notable example is a Soviet intelligence operation that propagated the false claim that the human immunodeficiency virus (HIV) and the resulting acquired immunodeficiency syndrome (AIDS) were biological weapons created by the United States [31]. Disinformation thus evolved into a structured geopolitical tool, used deliberately to influence public opinion and international relations.

In today's digital media landscape, the proliferation of fake news represents one of the most significant challenges to information integrity and democratic discourse. The rapid

dissemination of fake news through social media platforms and online information ecosystems has profoundly transformed how individuals consume, share, and interpret information.

Multiple studies have shown that social networks are highly effective at rapidly disseminating fake news to large audiences, thereby facilitating the manipulation of opinions and behaviours [9]. One key factor contributing to this effectiveness is the low cost and scalability of online influence operations. During the 2016 US presidential election, for instance, Russia's Internet Research Agency reportedly reached approximately 120 million Americans through social media advertisements costing about $100,000 [32]. This phenomenon has intensified over time: the number of states engaging in social media fake news campaigns increased from 28 in 2017 to 70 in 2019, nearly quadrupling in just two years [10].

The digital age has triggered an unprecedented explosion in the spread of false and misleading information through the rise of the Internet and social media platforms. A landmark study by Vosoughi et al. (2018) [6] revealed a disturbing reality: **false news is, on average, 70% more likely to be shared online than true news**. This phenomenon is partly explained by the greater novelty and emotional impact of false news, which attracts users' attention more effectively. The authors also found that **truth takes six times longer than false news to reach 1,500 people on social media**, highlighting the structural disadvantage faced by accurate information in digital environments.

The term **"fake news"** was popularized during the 2016 US presidential campaign. Due to its ambiguity and increasing politicization, the term is generally avoided in academic and institutional contexts, where it is replaced by the more precise concept of **disinformation**. According to the definition adopted by the European Digital Media Observatory (EDMO), **disinformation refers to information that satisfies the following four criteria**: it is verifiably

false or misleading, has the potential to cause harm to society, is disseminated intentionally, and aims to produce political or economic gain [33].

Fake news can erode trust in democratic institutions. Henschke et al. [32] show that electoral interference paves the way for discriminatory discourse, amplifying citizen distrust. While the 2016 U.S. elections marked a turning point, the phenomenon accelerates with every election cycle [34].

Beyond democratic processes, fake news propaganda can also serve military and geopolitical objectives [35]. Russia, for instance, has employed hybrid strategies that combine cyberattacks with coordinated information campaigns to weaken adversaries. Such tactics have been documented both in the context of the war in Ukraine and in influence activities targeting Canada, particularly for mitigating international economic sanctions [36].

The consequences of fake news can also be life-threatening, as evidenced during the COVID-19 pandemic. The World Health Organization (WHO) has identified vaccine hesitancy, driven in part by conspiracy theories, as one of the greatest threats to global public health [8]. In January 2021, huge fake news protests amplified through social media temporarily disrupted operations at the Dodger Stadium vaccination site. Furthermore, a Canadian study estimated that 2.35 million Canadians delayed or refused vaccination between March and November 2021, resulting in a 22% increase in COVID-19 infections and led to an additional 300 million dollars in healthcare costs [37].

Fake news dissemination can also be commercially driven. Clickbait-based fake news generates advertising revenue, sometimes at the risk of defamation litigation [38]. For instance, a toxic “grooming” narrative has been amplified by a small number of highly influential accounts, which the Center for Countering Digital Hate estimated could collectively generate up to USD 6.4 million annually in advertising revenue on Twitter [39].

The financial profitability of modern misinformation pipelines has led to an exponential increase in the volume and velocity of deceptive content online. Because manual fact-checking cannot scale to match this industrially driven propagation, computational methods have become indispensable for real-time identification and mitigation. To understand how automated systems intercept these sophisticated narratives, it is necessary to examine the underlying algorithmic frameworks designed to process textual data.

2.1. DEEP-LEARNING MODELS

Natural Language Processing (NLP) is the ability of a computer program to understand human language as it is spoken and written also called natural language. It is a component of artificial intelligence. One of the most important and difficult tasks in the overall NLP process is training a machine to derive the true meaning of words, especially when the same word can have several meanings within a single document. There are many words that share the same spelling but have different meanings.

NLP evolved from statistical and probabilistic methods widely used between the 1990s and early 2010s toward deep learning architectures, culminating in the introduction of Transformer-based models in 2017. It evolved from rule-based and statistical approaches toward deep neural architectures capable of learning contextual representations from large-scale corpora. Early NLP systems primarily relied on handcrafted linguistic rules and probabilistic models, whereas modern architectures leverage representation learning to capture semantic, syntactic, and contextual dependencies directly from data. In the context of fake news detection, this evolution has profoundly transformed the ability of models to identify deceptive linguistic patterns, contextual inconsistencies, and latent semantic cues.

Prior to the introduction of Transformers in 2017, we can identify two major algorithms in fake news detection: Recurrent Neural Network and Long Short-Term Memory.

2.1.1. RNN AND THE VANISHING GRADIENT PROBLEM

Recurrent Neural Networks (RNN) were introduced in the early 1990s, among the first neural architectures specifically designed to process sequential data such as text. Unlike feedforward neural networks, RNNs maintain a hidden state that is recursively updated at each time step, allowing information from previous tokens to influence subsequent predictions. Figure 2.1 illustrates the operating principle of an RNN.

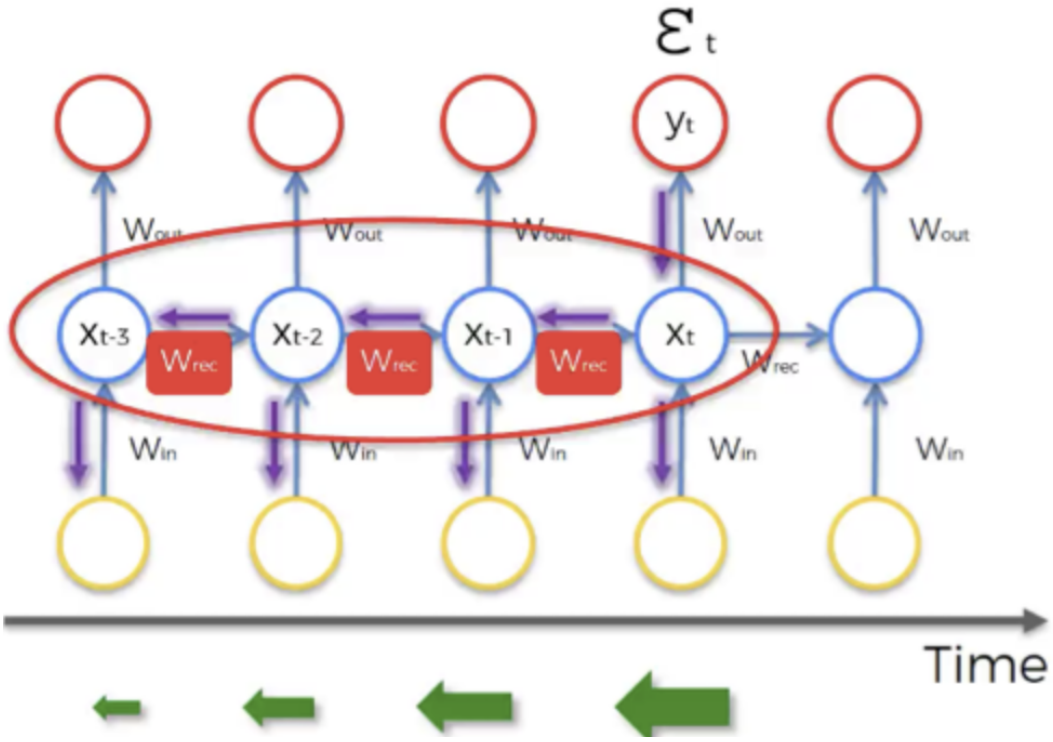


Figure 2.1 : Illustration of gradient propagation in a standard RNN

The hidden state of a standard RNN is defined as follows [40]:

$$h_t = \tanh(W_h h_{t-1} + W_x x_t + b) \tag{2.1}$$

The hidden state h_t is computed as a function of the current input x_t and the previous hidden state h_{t-1} where W_h and W_x are weight matrices, b is a bias term, and tanh is a non-linear activation function. This recursive formulation enables the network to encode temporal dependencies across a sequence.

However, training RNNs involves a fundamental optimization difficulty known as the *vanishing gradient problem* [41]. During Backpropagation Through Time, gradients are repeatedly multiplied by recurrent weights, which may lead to exponential decay (vanishing gradient) or growth (exploding gradient) depending on the spectral properties of the weight matrix W_h. The gradient of the loss $\mathscr{L}$ with respect to a hidden state h_t can be expressed as:

$$\frac{\partial \mathscr{L}}{\partial h_t} = \frac{\partial \mathscr{L}}{\partial h_T} \prod_{k=t}^{T-1} \frac{\partial h_{k+1}}{\partial h_k}. \tag{2.2}$$

When these spectral properties (eigenvalues) are smaller than one, successive multiplications cause the gradient to decay exponentially as it propagates backward through time. Consequently, early tokens in a long sequence receive an extremely weak learning signal, preventing the network from effectively capturing long-range dependencies. Conversely, eigenvalues greater than one may lead to exploding gradients, causing numerical instability. The phenomenon was first identified in foundational work by Hochreiter [41] and later formally analyzed by Bengio and colleagues [40].

In the context of textual analysis, this limitation is particularly critical. The semantic interpretation of a statement may depend on information occurring many tokens earlier, and

the inability to preserve long-range dependencies restricts the model's expressive capacity. This structural weakness motivated the development of gated recurrent architectures such as LSTM, specifically designed to regulate gradient flow and preserve information over extended contexts.

2.1.2. LSTM AND THE GATES

To address the vanishing gradient problem inherent to standard RNNs, Long Short-Term Memory (LSTM) networks were introduced in 1997 as a gated recurrent architecture specifically designed to regulate information flow across time steps. The architecture was first introduced in their seminal 1997 paper by Hochreiter and Schmidhuber [42]. Unlike conventional RNNs, which rely on a single hidden state, LSTMs incorporate an explicit memory cell c_t that enables the network to preserve information over extended temporal horizons. Figure 2.2 represents the architecture of a single LSTM cell.

The core innovation of LSTMs lies in their gating mechanisms, which control the addition, retention, and removal of information from the memory cell. At each time step t, the architecture computes three gates:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad \text{(forget gate)} \tag{2.3}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad \text{(input gate)} \tag{2.4}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad \text{(output gate)} \tag{2.5}$$

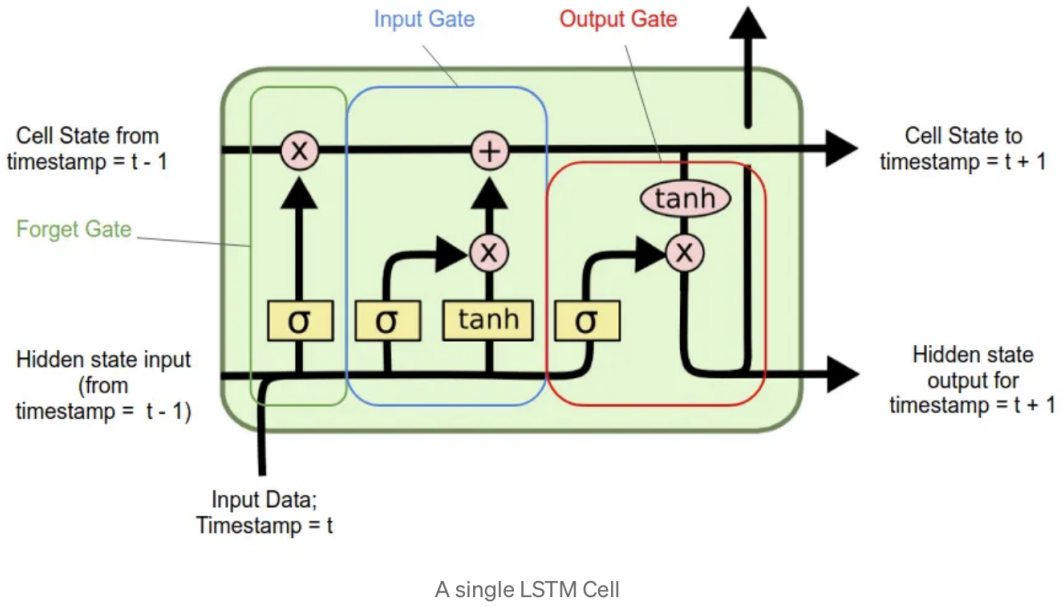


Figure 2.2 : Architecture of a single LSTM cell

where σ denotes the sigmoid activation function and $[h_{t-1}, x_t]$ represents the concatenation of the previous hidden state and the current input. The candidate memory update is computed as:

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c). \tag{2.6}$$

The memory cell is then updated according to:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{2.7}$$

where $\odot$ denotes element-wise multiplication. Finally, the hidden state is obtained as:

$$h_t = o_t \odot \tanh(c_t). \tag{2.8}$$

The additive structure of the memory update equation plays a crucial role in mitigating gradient decay. Because the previous cell state c_{t-1} is directly scaled and carried forward, gradients can propagate backward through time with reduced attenuation compared to standard RNNs. This mechanism enables LSTMs to capture longer-range dependencies in textual data and improves stability during training.

In practical applications, the behavior and performance of LSTM-based models are strongly influenced by a set of architectural and training parameters.

2.1.2.1. ARCHITECTURAL PARAMETERS

The core architecture of an LSTM network is defined by its structural configuration, which governs its capacity to model complex temporal relationships. A summary of these representative architectural parameters and their typical values is provided in Table 2.1.

- **Number of LSTM Units (Hidden Size)**: The number of LSTM units, also referred to as the hidden size, determines the dimensionality of the hidden state vector at each time step. This parameter directly controls the representational capacity of the model. A larger hidden size enables the LSTM to encode richer temporal and semantic patterns but increases computational cost and the risk of overfitting, particularly in limited-data settings. Conversely, a smaller hidden size constrains model capacity and may limit the ability to capture long-range dependencies in textual sequences.
- **Number of LSTM Layers (Depth)**: The number of stacked LSTM layers defines the depth of the recurrent network. In a multi-layer LSTM, the output of one layer serves as the input to the next, enabling the model to learn hierarchical temporal

representations. Deeper LSTM architectures can capture increasingly abstract sequential features; however, increasing depth also complicates training, may slow convergence, and requires careful regularization to avoid instability and overfitting.

- **Bidirectionality**: Bidirectional LSTMs process input sequences in both forward and backward temporal directions, concatenating representations from past-to-future and future-to-past contexts. This configuration allows the model to incorporate information from both preceding and succeeding tokens when encoding each position. Bidirectionality is particularly advantageous in text classification tasks, where the full sentence or document is available at inference time, but it increases memory usage and computational complexity.

- **Output Layer Configuration**: The final hidden state or sequence representation is mapped to task-specific predictions via an output layer. For binary classification tasks, this typically consists of a dense (fully connected) layer paired with a sigmoid activation function to output a probability distribution.

2.1.2.2. INPUT REPRESENTATION PARAMETERS

Before sequential data is processed by the recurrent layers, several parameters dictate how tokens are represented and structured.

- **Embedding Dimension**: The embedding dimension specifies the size of the dense vector representation associated with each token. Word embeddings encode semantic and syntactic information that serves as input to the LSTM. Higher-dimensional embeddings provide richer representations but increase model size and training complexity.

Embeddings may be initialized randomly or using pretrained vectors, depending on the experimental design.

- **Maximum Sequence Length**: LSTM models operate on sequences of fixed maximum length. This parameter defines the maximum number of tokens processed per input instance. Sequences longer than this limit are truncated, while shorter sequences are padded. The choice of maximum sequence length reflects a trade-off between preserving contextual information and maintaining computational efficiency.

- **Padding and Truncation Strategy**: Padding is applied to shorter sequences to achieve uniform length across batches, while truncation is applied to longer sequences. Common strategies include truncating from the end or the beginning of the sequence. Padding masks are typically used to prevent padded tokens from influencing the hidden state updates during training.

- **Vocabulary Representation**: The vocabulary representation defines the mapping mechanism that translates textual tokens into continuous vector spaces. This can be achieved through discrete word-level indexing initialized randomly and learned dynamically during training, or by leveraging static, pretrained semantic word embeddings (such as Word2Vec or GloVe). This representation establishes the semantic foundation upon which the LSTM sequences its temporal transitions.

2.1.2.3. TRAINING AND OPTIMIZATION PARAMETERS

The training and optimization process of an LSTM network is governed by a set of hyperparameters that control how the model learns from sequential data. These parameters directly influence convergence stability, training efficiency, and the model's ultimate generalization capability.

- **Learning Rate**: controls the magnitude of weight updates during training. A learning rate that is too high may lead to unstable training and divergence, while a learning rate that is too low can slow convergence or result in suboptimal solutions. Learning rate selection is particularly critical for recurrent architectures, which are sensitive to optimization dynamics.

- **Optimizer**: determines how gradients are used to update model parameters. Adaptive optimizers such as Adam are commonly employed in LSTM training due to their ability to adjust learning rates dynamically for each parameter. The choice of optimizer affects convergence speed, stability, and generalization performance.

- **Batch Size**: defines the number of training instances processed simultaneously before updating model parameters. Smaller batch sizes introduce more gradient noise, which can improve generalization but slow training, whereas larger batch sizes offer computational efficiency at the risk of reduced generalization.

- **Number of Epochs**: specifies how many times the model iterates over the entire training dataset. Insufficient epochs may lead to underfitting, while excessive epochs increase the risk of overfitting. Epoch selection is typically guided by validation performance and early stopping criteria.

- **Loss Function**: defines the mathematical objective that the optimizer minimizes during the training phase. For sequence classification tasks focusing on binary targets, binary cross-entropy is standard, as it measures the distance between the model's predicted probability distributions and the actual binary ground-truth labels.

2.1.2.4. REGULARIZATION PARAMETERS

To mitigate the risk of overfitting, especially when dealing with complex recurrent architectures and high-dimensional textual features, specific regularization mechanisms are integrated into the LSTM framework. These parameters are essential for ensuring that the model generalizes well to unseen data rather than merely memorizing the training sequences.

- **Dropout**: is a regularization technique that randomly deactivates a fraction of neurons during training. In LSTM networks, dropout is commonly applied to input-to-hidden and hidden-to-output connections to reduce overfitting. The dropout rate determines the proportion of units that are temporarily removed during each training iteration.
- **Recurrent Dropout**: applies dropout to the recurrent connections within the LSTM cell. Unlike standard dropout, recurrent dropout uses the same dropout mask across time steps, preserving temporal consistency. This technique can improve generalization but may negatively impact learning if applied too aggressively.

From an inductive bias perspective, LSTMs remain fundamentally sequential models: information is processed token by token, and the hidden state evolves stepwise along the sequence. While gating improves memory retention, the architecture still exhibits limitations when modeling very long contexts, as information must be compressed into a fixed-size state vector [43]. Moreover, the inherently sequential computation prevents full parallelization during training, making LSTMs computationally less efficient than attention-based architectures.

In the context of fake news detection, LSTMs are particularly suited for capturing stylistic patterns, sequential rhetorical structures, and local contextual dependencies. However, their constrained memory representation and sequential bottleneck (parallelization) limit their

Table 2.1 : Representative LSTM Parameters

Parameters	Role and Impact	Typical Values
Architecture		
Hidden size	Defines the dimensionality of the hidden state and memory cell, controlling temporal modeling capacity	64, 128, 256
Number of LSTM layers	Determines network depth via stacked recurrent layers	1–3 layers
Bidirectionality	Processes sequences in one or both temporal directions to capture context	Unidirectional, Bidirectional
Output layer configuration	Maps sequence representations to task-specific predictions	Dense layer + sigmoid activation
Input Representation and Sequence Handling		
Embedding dimension	Dimensionality of input token embeddings fed to the LSTM	100, 200, 300
Maximum sequence length	Maximum number of tokens processed per input sequence	100–500 tokens
Padding and truncation	Standardizes sequence lengths for batch processing	Post-padding, pre/post truncation
Vocabulary representation	Defines how tokens are mapped to embeddings	Word-level or pretrained embeddings
Training and Optimization		
Learning rate	Controls the magnitude of parameter updates during training	10^{-4}–10^{-3}
Optimizer	Gradient-based optimization algorithm	Adam, RMSprop
Batch size	Trade-off between convergence stability and training efficiency	16–128
Number of epochs	Number of full passes over the training dataset	5–50
Loss function	Task-specific optimization objective	Binary cross-entropy
Regularization		
Dropout rate	Regularization applied to non-recurrent connections	0.2–0.5
Recurrent dropout	Dropout applied to recurrent connections to improve generalization	0.1–0.3

ability to model complex global interactions across distant parts of a document [44]. These architectural constraints partly explain the paradigm shift toward Transformer-based models, which replace recurrence with attention mechanisms capable of modeling global dependencies more directly.

2.1.3. TRANSFORMER: FROM RECURRENCE TO ATTENTION

Although LSTMs significantly mitigate the vanishing gradient problem and improve the modeling of long-range dependencies, they remain constrained by a fundamental architectural limitation: sequential computation. Each hidden state depends on the previous one, which imposes an intrinsic temporal bottleneck. This sequential dependency not only limits parallelization during training but also forces the model to compress all prior contextual information into a fixed-size state vector. As sequence length increases, this compression becomes lossy, reducing the model's ability to represent complex global interactions within text.

The Transformer architecture, introduced by Vaswani et al. [44] in 2017, represents a structural departure from recurrence by replacing sequential processing with self-attention mechanisms. Instead of propagating information stepwise across time, Transformers allow each token in a sequence to directly attend to all other tokens simultaneously. This global receptive field eliminates the need for iterative state propagation and enables fully parallel computation.

At the core of the Transformer lies the self-attention mechanism. For a sequence represented by input embeddings X, three matrices are computed: queries (Q), keys (K), and values (V). Attention weights are obtained by scaled dot-product attention:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.9}$$

where d_k denotes the dimensionality of the key vectors. This operation allows each token to dynamically weight the importance of every other token in the sequence. Consequently, contextual dependencies are modeled explicitly rather than implicitly stored in a recurrent hidden state.

From an inductive bias perspective, Transformers shift the modeling assumption from temporal continuity to relational connectivity. Dependencies are no longer constrained by distance in the sequence; instead, they are determined by learned attention weights. This structural property is particularly advantageous for textual analysis tasks such as fake news detection, where critical cues may be distributed across distant parts of a document or embedded in subtle semantic inconsistencies.

By removing the sequential bottleneck and enabling global context modeling, Transformers overcome the primary limitations of recurrent architectures. This paradigm shift laid the foundation for encoder-based models such as BERT and, subsequently, LLMs such as Gemini, Copilot, GP-4, LLaMA2, which rely on large-scale pretraining to learn rich contextual representations. The transition from recurrence to attention thus constitutes not merely an incremental improvement, but a fundamental reconfiguration of how linguistic information is represented and processed.

The Transformer framework can be instantiated in three principal architectural configurations: encoder-only models, decoder-only models, and encoder–decoder models. While all three rely on self-attention mechanisms, they differ fundamentally in training objectives, inductive biases, and downstream applications. Understanding these distinctions is essential for positioning models such as BERT and LLM within the broader landscape of natural language processing as shown in Figure 2.3.

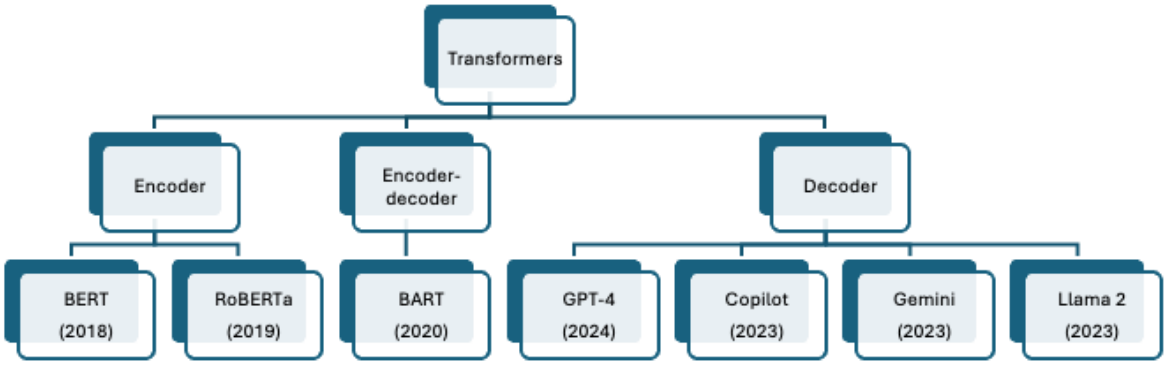


Figure 2.3 : Comparison between Transformers

Encoder-only models are designed to produce contextualized representations of an input sequence. Their primary objective is representation learning rather than sequence generation. During pretraining, these models typically rely on masked language modeling (MLM), in which a subset of tokens is masked and the model learns to predict them using bidirectional context. Because attention is computed across the entire sequence without causal constraints, encoder-only models capture rich bidirectional dependencies. This inductive bias makes them particularly well suited for discriminative tasks such as text classification, sentiment analysis, and fake news detection, where the goal is to assign a label to an existing text rather than generate new content. BERT is the canonical example of this architecture.

Decoder-only models are autoregressive architectures trained to predict the next token in a sequence given all previous tokens. Unlike encoder-only models, they employ causal masked self-attention, which prevents each token from attending to future tokens. The training objective maximizes the likelihood of generating the next token in a sequence. This autoregressive formulation induces a generative inductive bias: the model learns to produce coherent text by modeling sequential probability distributions. LLMs, such as GPT-style architectures, belong to this category. Their strength lies in text generation, reasoning through token-by-token continuation, and flexible prompting. While decoder-only models can perform

classification tasks via prompting, their architecture is fundamentally optimized for generation rather than discriminative representation learning.

Encoder–decoder models combine both components within a single architecture. The encoder processes the input sequence into contextual representations, and the decoder generates an output sequence conditioned on these representations through cross-attention mechanisms. This configuration is particularly effective for sequence-to-sequence tasks such as machine translation, summarization, and text rewriting. During training, the decoder attends both to previously generated tokens (via masked self-attention) and to encoder outputs (via cross-attention), enabling structured transformation of input text into output text. Models such as BART and T5 exemplify this architecture.

From the perspective of fake news detection, these architectural distinctions have direct methodological implications. Encoder-only models are naturally suited for classification tasks due to their bidirectional contextual representations. Decoder-only models, while generative, can leverage large-scale pretraining to perform zero-shot or few-shot detection through implicit reasoning. Encoder–decoder architectures occupy an intermediate position, enabling claim reformulation, summarization, or fact-checking-style transformations. Thus, the evolution from encoder-based discriminative models to generative large-scale decoders reflects not merely an increase in model size, but a shift in modeling objectives from representation learning to probabilistic text generation. This distinction is central to understanding the methodological choices made in this dissertation, particularly the comparison between BERT-based classification approaches and LLM-based reasoning approaches.

2.1.4. BERT

BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. in 2018 [45], represents a major breakthrough in natural language processing. Its core innovation lies in its pretraining strategy, particularly the Masked Language Modeling (MLM) objective. During pretraining, a subset of tokens in the input sequence is randomly masked, and the model is trained to predict these masked tokens using both left and right context. Formally, given an input sequence $X = (x_1, x_2, ..., x_n)$, the MLM objective maximizes:

$$\mathscr{L}_{MLM} = -\sum_{i \in M} \log P(x_i \mid X_{\setminus M}), \tag{2.10}$$

where M denotes the set of masked positions and $X_{\setminus M}$ represents the unmasked tokens. This bidirectional conditioning enables BERT to learn deep contextual representations that capture semantic and syntactic dependencies more effectively than unidirectional models.

Architecturally, BERT consists of stacked Transformer encoder layers based on the self-attention mechanism introduced in Section 2.1.3. Multi-head self-attention enables the model to compute contextual interactions between all token pairs simultaneously, allowing the capture of syntactic, semantic, and discourse-level dependencies. Figure 2.4 presents a simplified BERT architecture highlighting its main components.

In contrast to LSTM-based models, BERT relies on a Transformer encoder architecture and benefits from large-scale pretraining on unlabeled corpora. As a result, only a subset of parameters is explicitly optimized during task-specific fine-tuning. For consistency with the LSTM configuration, BERT parameters are grouped into architectural, input representation, fine-tuning, and regularization parameters.

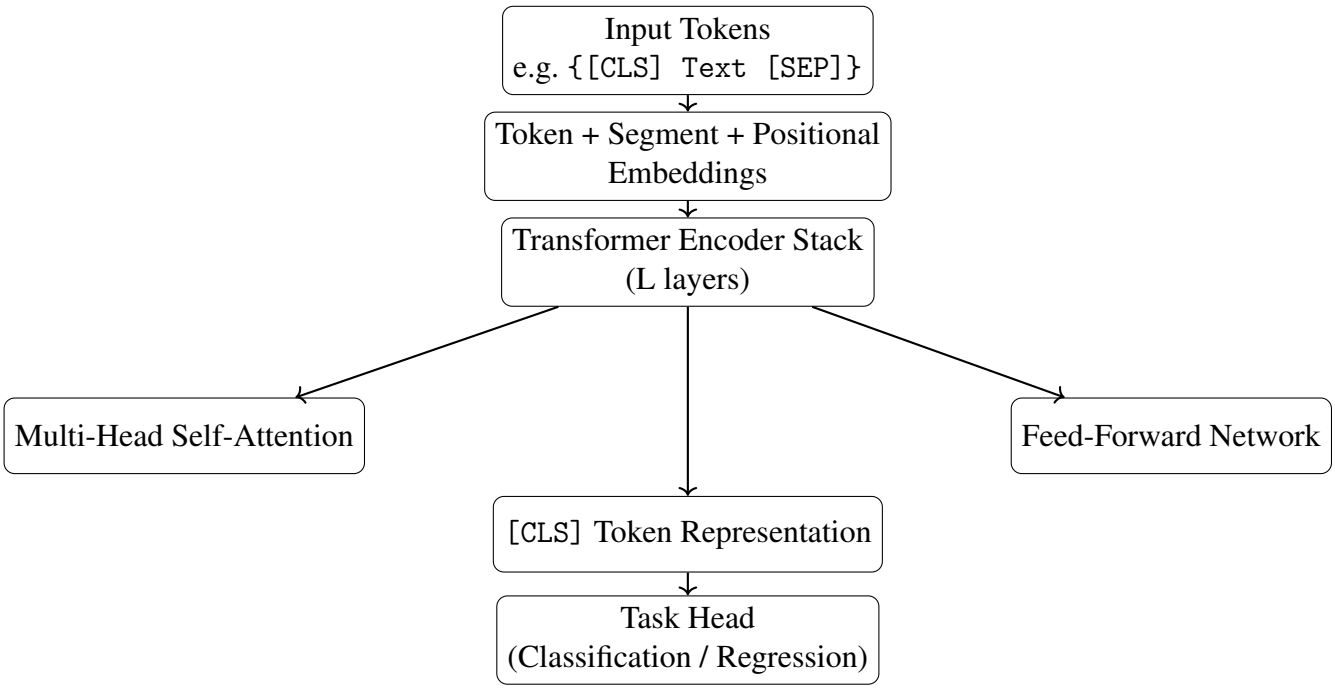


Figure 2.4 : Architecture of BERT

2.1.4.1. ARCHITECTURAL PARAMETERS

The foundational architecture of the BERT model relies on stacked Transformer blocks that leverage self-attention mechanisms to learn deep, bidirectional contextual representations. These core structural parameters define the model's depth, width, and parallel processing capacity.

- **Number of Transformer Layers (Depth)**: defines the depth of the BERT model and determines the hierarchy of contextual representations learned across layers. Each layer applies self-attention and feed-forward transformations to refine token representations. Deeper models are capable of capturing increasingly abstract linguistic patterns but incur higher computational cost and memory usage. Standard pretrained configurations

(e.g., BERT-Base or BERT-Large) are employed, where the number of layers is fixed and inherited from the pretrained architecture.

- **Hidden Size**: specifies the dimensionality of the contextualized token embeddings produced at each layer of the Transformer. This parameter directly controls the representational capacity of the model. Larger hidden sizes allow richer semantic encoding and more expressive contextual interactions among tokens. Unlike LSTM hidden states, which evolve sequentially over time, BERT hidden representations are computed simultaneously for all tokens using self-attention mechanisms.

- **Number of Attention Heads**: determines how many parallel self-attention subspaces are used within each Transformer layer. Each attention head learns to focus on different aspects of contextual relationships, such as syntactic dependencies or long-distance semantic interactions. Increasing the number of heads improves the model's ability to capture diverse contextual patterns but increases computational complexity. The number of heads is fixed in pretrained BERT models and aligned with the hidden size.

- **Maximum Sequence Length**: Due to the quadratic computational and memory complexity of the self-attention mechanism, BERT enforces an architectural limit on input sequence length. Typically constrained to a maximum of 512 tokens, this parameter requires longer documents to be truncated or segmented during processing.

2.1.4.2. INPUT REPRESENTATION PARAMETERS

Before entering the Transformer layers, textual data must be transformed into a structured format that captures token identities, order, and overall context. These parameters dictate how raw text is tokenized, vectorized, and enriched with spatial awareness based on large-scale textual knowledge.

- **Token Embedding and Subword Tokenization**: BERT employs a subword-based tokenization scheme, typically WordPiece, to represent input text. This approach decomposes words into smaller units, allowing the model to handle rare and out-of-vocabulary terms effectively. Token embeddings are combined with segment embeddings and positional embeddings to form the input representation for each token. These embeddings are pretrained on large corpora and provide rich lexical and semantic information prior to fine-tuning.

- **Positional Embeddings**: Since the Transformer architecture lacks an inherent notion of sequence order, positional embeddings are added to token embeddings to encode the relative and absolute positions of tokens within a sequence. These embeddings enable the model to distinguish between different word orders and maintain syntactic coherence. Positional embeddings in BERT are learned during pretraining and impose a fixed maximum sequence length.

- **Pretraining Corpora**: The foundational linguistic knowledge embedded within BERT is acquired by training on massive, unlabeled text datasets. The standard pretraining pipeline leverages extensive corpora such as English Wikipedia and BooksCorpus, allowing the model to develop a robust baseline understanding of syntax, grammar, and world knowledge prior to task-specific adaptation.

2.1.4.3. PRETRAINING PARAMETERS

The primary strength of BERT lies in its self-supervised pretraining phase, where the model learns bidirectional language representations without human labeling. This process is driven by two simultaneous learning objectives which force the network to understand both internal token context and macro-level sentence relationships.

- **Masked Language Modeling (MLM)**: To overcome the limitation of traditional unidirectional language models, the MLM objective randomly masks a specific percentage (typically 15%) of input tokens. The model is then tasked with predicting the original identities of these corrupted tokens using only the surrounding bidirectional context, forcing the layers to learn deep, interrelated representations.

- **Next Sentence Prediction (NSP)**: To capture inter-sentence coherence and long-range discourse relations, the NSP objective frames pretraining as a binary classification task. The model is presented with pairs of sentences and must determine whether the second sentence logically follows the first in the original text, enhancing its capacity for downstream tasks like question answering or text entailment.

- **Pretraining Strategy**: The pretraining strategy governs how the model is exposed to the large-scale corpora over multiple training iterations. This phase involves optimization over several epochs with massive batch sizes to ensure stable representation learning and optimal parameter generalization before any fine-tuning takes place.

2.1.4.4. FINE-TUNING PARAMETERS

Fine-tuning adapts the pretrained, general-purpose BERT model to specific downstream applications by updating its weights on a smaller, task-specific labeled dataset. This optimization process requires specialized hyperparameter configurations to prevent the catastrophic forgetting of pretrained knowledge while maximizing classification performance.

- **Learning Rate**: controls the magnitude of parameter updates during fine-tuning. Due to BERT's pretrained nature, fine-tuning requires significantly smaller learning rates than models trained from scratch. Learning rates that are too high may lead to catastrophic

forgetting of pretrained knowledge, while overly small learning rates may hinder effective task adaptation. Consequently, careful learning rate selection is critical for stable and effective fine-tuning.

- **Optimizer**: Fine-tuning of BERT models typically employs the AdamW optimizer, which extends the Adam algorithm by decoupling weight decay from gradient updates. This optimization strategy improves generalization performance and training stability, particularly in large Transformer-based models. AdamW is well-suited to fine-tuning scenarios where small learning rates and regularization are essential.

- **Batch Size**: determines the number of training samples processed before parameter updates. In BERT fine-tuning, batch size is often constrained by GPU memory requirements due to the model's large number of parameters and the quadratic complexity of self-attention. Smaller batch sizes may improve generalization but increase training time, while larger batch sizes enhance computational efficiency at the cost of higher memory usage.

- **Number of Epochs**: specifies how many passes over the task-specific training data are performed. Because BERT representations are already highly informative, only a small number of epochs is typically required. Excessive fine-tuning may result in overfitting, particularly on small datasets, making careful monitoring of validation performance essential.

- **Loss Function**: The optimization objective during fine-tuning is dictated by a task-specific loss function. For classification tasks, cross-entropy loss is applied to quantify the error between the model's predicted probability distributions and the actual ground-truth labels.

- **Pooling Strategy**: To perform sequence-level classification, the token-level outputs must be aggregated into a single vector representation. BERT's standard pooling strategy extracts the final hidden state of the special `[CLS]` token, which is positioned at the start of every input sequence and serves as an encoded summary of the entire sentence or document.

2.1.4.5. REGULARIZATION PARAMETERS

Due to the immense parameter size of the BERT architecture, strict regularization is required during both pretraining and fine-tuning to prevent overfitting. These mechanisms focus on introducing stochastic noise within the hidden layers and the self-attention matrices to ensure robust feature extraction.

- **Dropout**: is applied to embedding layers, hidden representations, and output layers during fine-tuning to reduce overfitting. The dropout rate controls the fraction of units that are randomly deactivated during training iterations. Moderate dropout values are generally sufficient due to the strong regularization effect already provided by pretraining.
- **Attention Dropout**: is applied specifically to the attention weight matrices within the self-attention mechanism. This form of regularization prevents the model from relying too heavily on specific attention patterns and promotes more robust contextual learning. Attention dropout serves a role analogous to recurrent dropout in LSTM networks, regularizing the core dependency modeling mechanism.

BERT's input representation is constructed by summing three types of embeddings: token embeddings, positional embeddings, and segment embeddings. Token embeddings

encode the identity of each subword unit, positional embeddings preserve word order, and segment embeddings distinguish between paired sentences in tasks such as next-sentence prediction. These embeddings are then processed by the Transformer encoder layers to produce contextualized token representations.

Table 2.2 provides a detailed overview of the main architectural, pretraining, and fine-tuning parameters associated with BERT.

Following BERT's success, several improved variants were proposed to address its architectural and computational limitations. RoBERTa shows that removing the next-sentence prediction objective, increasing training data, and optimizing hyperparameters significantly improved performance across multiple benchmarks. ALBERT introduced parameter sharing and factorized embeddings to reduce memory consumption while maintaining competitive accuracy [46]. These models illustrated that architectural and training optimizations could improve both efficiency and representational quality.

Another line of research extended BERT to multilingual and cross-lingual applications. Cross-Lingual Language Model-RoBERTa (XLM-R, trained on two terabytes of multilingual text, enabled robust performance across many languages without requiring language-specific training data [47]. This development is particularly relevant for fake news detection, where deceptive content often appears across multiple linguistic contexts and social media environments
.

2.1.4.6. TOPIC-CLUSTERING MODELS

While BERT processes each message individually, fake news rarely operates at the level of isolated posts. Instead, it spreads through recurrent narratives, thematic patterns,

Table 2.2 : Unified BERT Parameters

Parameter	Role and Impact	Typical Values
Architecture		
Number of Transformer layers	Controls model depth and hierarchical representation learning	12 (Base), 24 (Large)
Hidden size	Dimensionality of contextualized token representations	768 (Base), 1024 (Large)
Number of attention heads	Enables parallel self-attention over multiple representation subspaces	12 (Base), 16 (Large)
Maximum sequence length	Architectural limit on input length due to attention complexity	128, 256, 512
Input Representation		
Embedding type	Subword token representation enabling open-vocabulary modeling	WordPiece
Positional embeddings	Explicitly encode token order information	Learned positional embeddings
Pretraining corpora	Large-scale unlabeled text used to learn general language representations	Wikipedia, BooksCorpus
Pretraining Objectives		
Masked Language Modeling (MLM)	Learns bidirectional contextual token representations	15% masked tokens
Next Sentence Prediction (NSP)	Encodes inter-sentence coherence and discourse relations	Binary classification
Pretraining strategy	Defines exposure and generalization during representation learning	Multiple epochs over corpus
Fine-Tuning		
Learning rate	Controls magnitude of updates during task adaptation	2×10^{-5}–5×10^{-5}
Optimizer	Optimization algorithm used during fine-tuning	AdamW
Batch size	Stability–efficiency trade-off under memory constraints	16–32
Number of epochs	Controls degree of adaptation to downstream task	2–4
Loss function	Task-specific optimization objective	Cross-entropy
Pooling strategy	Sequence-level representation used for downstream tasks	`[CLS]` token
Regularization		
Dropout rate	Regularization of hidden representations	0.1
Attention dropout	Regularization applied to attention weights	0.1

and discursive frames that reappear across hundreds or even thousands of messages. In the COVID-19 dataset, for example, many posts revolve around a limited set of recurring misinformation themes such as "vaccines cause infertility," "the pandemic was planned," "5G spreads the virus," or "pharmaceutical companies are hiding the truth." These narratives form the backbone of coordinated disinformation campaigns: they persist over time, are repeated by different users, and are adapted across platforms. Detecting fake news therefore requires not only classifying individual posts but also identifying the broader narrative structures within which these messages circulate.

While Transformer-based classifiers such as BERT excel at evaluating individual posts, fake news rarely spreads as isolated messages. Instead, it propagates through recurring narratives and thematic patterns that reappear across large volumes of content. Classifying posts one by one therefore provides only a micro-level view of the phenomenon. To capture the broader structure of fake news campaigns, it is necessary to identify clusters of messages that revolve around the same topic, rhetorical frame, or conspiracy narrative. Topic-clustering models such as BERTopic and Top2Vec address this need by grouping semantically similar posts into coherent themes, enabling the detection of dominant misinformation narratives, emerging topics, and coordinated patterns of discourse. Integrating clustering methods alongside BERT thus provides a complementary macro-level perspective on how fake news circulates and evolves. BERTopic combines Transformer-based embeddings with class-based TF-IDF to generate interpretable topic representations, making it well suited for identifying fine-grained misinformation themes. Top2Vec, in contrast, jointly learns document and topic embeddings in a shared vector space, enabling the discovery of topics without requiring predefined labels or manual preprocessing.

Applied to fake news datasets, these clustering models reveal the dominant narratives circulating within the data, highlight emerging or rapidly evolving themes, and expose co-

ordinated messaging patterns. For example, clusters may correspond to conspiracy theories, anti-vaccine rhetoric, political disinformation, or pseudoscientific claims. By analyzing these clusters, researchers can track how narratives evolve over time, how they interact with real-world events, and how they are amplified by specific user groups or automated accounts.

Narrative-level detection thus complements post-level classification: while BERT identifies whether a specific message is likely to be fake, clustering models uncover the broader thematic context in which fake news operates. Together, these approaches provide a more comprehensive understanding of the structure, dynamics, and propagation mechanisms of online fake news.

Clustering models such as BERTopic and Top2Vec directly address this phenomenon by grouping semantically similar messages into coherent topics, allowing researchers to identify the dominant misinformation narratives that propagate through user networks. Whereas profile-based, network-based, and temporal analyses capture how misinformation spreads, clustering reveals what is being spread at the narrative level. These models therefore provide a natural extension to context-based detection by uncovering the thematic structures that underlie coordinated disinformation campaigns.

2.1.4.7. ARCHITECTURAL ENHANCEMENTS

More recently, ModernBERT (2024) introduced several architectural enhancements, including Rotary Positional Embeddings (RoPE), FlashAttention, and extended context windows [48]. RoPE improves extrapolation to longer sequences, enabling the model to process news articles far beyond the original 512-token limit. FlashAttention reduces memory complexity and accelerates inference, making encoder-only architectures more scalable. Although ModernBERT has not yet been applied specifically to fake news detection, peer-reviewed

evaluations show superior performance across a range of natural language understanding tasks, suggesting strong potential for future applications.

Despite these advances, BERT remains a discriminative encoder-only model. Its self-attention weights do not provide faithful explanations of model decisions, as shown by Jain and Wallace [49], and its fixed input length limits its ability to process long articles without truncation. These limitations motivate the integration of explainability methods and contextual metadata, which are explored in the subsequent chapters of this dissertation.

In summary, BERT assumes that meaning emerges from global contextual interaction rather than sequential accumulation. This property is particularly advantageous for fake news detection, where credibility cues may depend on subtle semantic inconsistencies, rhetorical exaggeration, or context-dependent framing. Bidirectional contextualization enables the detection of misleading constructions even when individual words appear factually neutral, making BERT a foundational component of modern fake news detection systems.

While Transformer-based architectures such as BERT focus on contextual semantic understanding at the textual level, other approaches attempt to analyze fake news through thematic propagation patterns and discourse clustering across large-scale corpora.

2.1.5. EXPLAINABLE AI

As deep learning models have grown in complexity, their decision-making processes have become increasingly opaque. While Transformer-based architectures such as BERT achieve high predictive performance, their internal representations are distributed across millions of parameters, rendering their reasoning processes difficult to interpret. This opacity is commonly referred to as the black box problem.

Explainable AI (XAI) seeks to mitigate this limitation by providing insights into the features and mechanisms driving model predictions. XAI, is a research field that investigates how AI decisions and the data driving those decisions can be made transparent and understandable to individuals [50]. This transparency may help users critically evaluate automated decisions, including building trust with users and regulatory requirements [51].

XAI, as defined in NISTIR 8312, is built on four foundational principles that ensure AI systems can justify and communicate their decisions in a trustworthy way. First, an explainable system must provide evidence or reasons for its outputs, allowing users to understand why a particular decision was made. Second, these explanations must be meaningful, meaning they are tailored to the knowledge and needs of different users. Third, explanations must be accurate, faithfully reflecting the system's true internal processes rather than simplified or misleading narratives. Finally, AI systems must respect their knowledge limits, operating only under conditions for which they were designed and producing outputs only when sufficiently confident. Together, these principles form a multidisciplinary framework that supports transparency, reliability, and user trust in AI systems.

XAI provides significant benefits in various areas: fostering trust and confidence, enhancing decision-making, facilitating debugging and improvement, and ensuring regulatory compliance [52]. For instance, under the GDPR, individuals are entitled to meaningful information about automated decisions affecting them. However, XAI presents several challenges. One persistent issue is measuring the effectiveness and faithfulness of explanations [53].

Within the NLP domain, XAI methods for Transformer architectures are commonly categorized into four major classes: post-hoc model-agnostic methods, gradient-based attribution techniques, Transformer-specific explainability approaches, and occlusion-based analyses. All

categories are indeed post-hoc (no backpropagation), but they differ in model-agnosticism, gradient dependence, and architecture specificity.

Post-hoc model-agnostic methods, such as SHAP and LIME, aim to explain predictions without modifying the underlying model. These methods perturb the input text (e.g., by masking or removing words) and observe the resulting changes in the predicted output. For example, if removing the word *"breaking"* significantly decreases the probability of fake news, then this word is considered highly influential. These approaches are intuitive and produce human-interpretable explanations at the word level.

Gradient-based methods (e.g., Integrated Gradients or Gradient$\times$Input) rely on the differentiable nature of neural networks. They compute how sensitive the model's prediction is to small changes in each input feature. In the context of text, this means assigning an importance score to each token based on how much it contributes to the prediction. These methods are more faithful to the internal model structure but less intuitive to explain.

Transformer-specific approaches focus on the attention mechanism unique to models like BERT. Techniques such as attention visualization, attention rollout, or layer-wise relevance propagation (LRP) aim to trace how information flows across layers and attention heads. For instance, they can show which words the model attends to when processing a specific token, helping to visualize internal dependencies within the sentence.

Occlusion-based methods provide a more direct way to measure causal importance. They systematically remove or mask parts of the input (such as words or phrases) and observe how the prediction changes. If removing a word drastically alters the output, that word is considered essential for the decision.

These four families of methods are complementary. Substantial empirical evidence shows that attention alone does not guarantee faithful or causally grounded explanations of model decisions. While attention reflects intermediate relational weighting during forward propagation, it does not necessarily correspond to the features that causally drive the final output [54]. Therefore, model-agnostic approaches provide intuitive explanations, gradient-based methods ensure faithfulness to the model, Transformer-specific techniques reveal internal mechanisms, and occlusion methods test causal importance. Together, they provide a comprehensive toolkit for explaining Transformer-based models such as BERT in high-stakes applications like fake news detection.

Figure 2.5 provides an overview of the main families of explainability methods used in Transformer-based models.

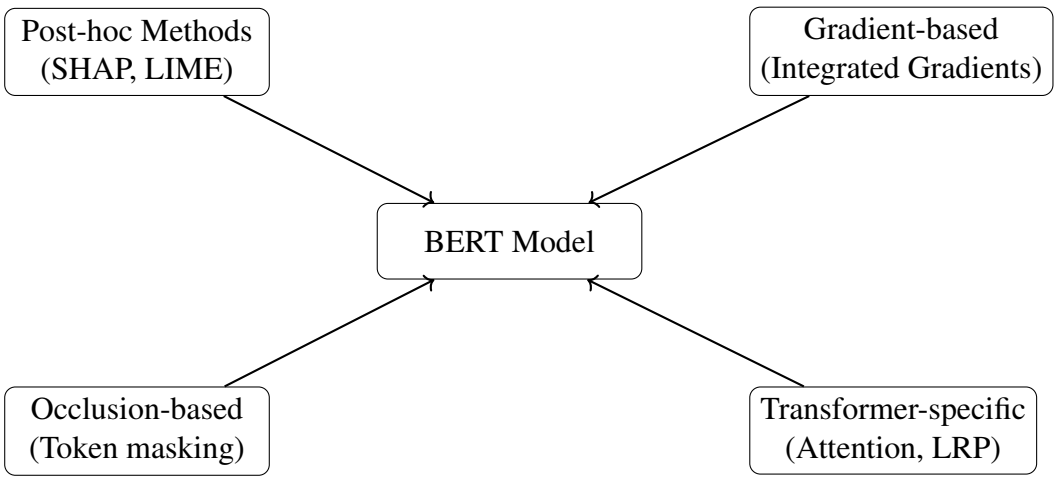


Figure 2.5 : Main categories of XAI methods applied to Transformer-based models such as BERT.

2.1.5.1. SHAPLEY METHOD

SHAP (2017) assigns each feature a contribution score based on Shapley values from cooperative game theory [55]. The Shapley value of feature i is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right], \tag{2.11}$$

where F is the set of all features and $f(S)$ denotes the model output given feature subset S.

where:

- F denotes the full set of features,
- S is a subset of features not containing i,
- $f(S)$ represents the model output when only features in S are present.

The unique appeal of SHAP lies in its strict adherence to game-theoretic axioms, which guarantees that the feature attributions are mathematically consistent. Specifically, SHAP satisfies three foundational properties:

1. **Local Accuracy (Efficiency):** The sum of the feature attributions must equal the difference between the model's prediction for the specific instance x and the expected baseline prediction:
$$f(x) = \phi_0 + \sum_{i \in F} \phi_i \tag{2.12}$$
where $\phi_0 = \mathbb{E}[f(x)]$ represents the base value (the average prediction of the model over a background dataset).

2. **Missingness:** If a feature i is absent from the input, its assigned contribution must be zero:

$$x_i = 0 \implies \phi_i = 0 \tag{2.13}$$

In text classification, this ensures that words or tokens not present in the evaluated sentence do not receive an arbitrary attribution score.

3. **Consistency (Monotonicity):** If a model changes such that the marginal contribution of feature i increases or stays the same for all coalitions, its Shapley value cannot decrease:

$$f'_x(S \cup \{i\}) - f'_x(S) \geq f_x(S \cup \{i\}) - f_x(S) \implies \phi_i(f') \geq \phi_i(f) \tag{2.14}$$

In Natural Language Processing (NLP) applications, evaluating $f(S)$ poses a non-trivial challenge. Standard architectures like BERT require dense contextual representations and are structurally unequipped to process "missing" words without breaking the syntactic alignment. To approximate the absence of a token, practitioners typically rely on conditional expectations, which can be computationally approximated using KernelSHAP.

2.1.5.2. LIME METHOD

LIME (2016) approximates the complex model f locally around an instance x by fitting a simple explainable model g through weighted perturbations:

$$\xi(x) = \underset{g \in G}{\arg\min}\, \mathscr{L}(f, g, \pi_x) + \Omega(g) \tag{2.15}$$

- f : The complex model (e.g., ModernBERT) that outputs a probability (e.g., $P(\text{Fake})$).

- $g \in G$: An interpretable model (e.g., a linear regression $g(z) = w \cdot z$) from the set of potential simple models G.

- π_x : The proximity kernel, which weights samples z based on how close they are to the original input x:

$$\pi_x(z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right) \tag{2.16}$$

Unlike SHAP, LIME does not guarantee consistency and may exhibit instability depending on the sampling strategy and kernel width. However, it remains computationally more efficient and flexible. Table 2.3 summarizes the key theoretical and practical differences between SHAP and LIME in NLP applications.

Table 2.3 : Comparison of SHAP and LIME in NLP contexts

Property	SHAP	LIME
Theoretical grounding	Game theory	Local surrogate
Consistency guarantee	Yes	No
Computational cost	High	Moderate
Stability	Relatively high	Sampling-dependent
Model-agnostic	Yes	Yes

Explainability has become increasingly important due to the rapid transformation of the modern information ecosystem. The threat landscape has evolved from isolated cases of disinformation to industrial-scale information manipulation campaigns. The convergence of state-sponsored information operations, commonly referred to as Foreign Information Manipulation and Interference (FIMI), and generative artificial intelligence now enables the large-scale production of highly realistic synthetic content, including deepfakes and persuasive textual narratives, at negligible cost. This industrialization of content generation contributes

to an information saturation effect in which misleading and legitimate information become increasingly difficult to distinguish.

At the same time, the adoption of artificial intelligence systems has significantly increased across society. While early deep learning models were often criticized for their lack of transparency, recent studies indicate a growing pragmatic acceptance of AI systems despite their inherent opacity. For instance, user trust in AI-enabled services has evolved from initial skepticism in 2023 to broader adoption in 2025, driven by perceived benefits such as efficiency and personalization.

Despite this increased adoption, the demand for transparency has not diminished. Instead, it has shifted from an ethical consideration to a regulatory requirement, particularly under frameworks such as the European AI Act. In this context, explainability becomes a necessary component of trustworthy AI systems rather than an optional feature.

To address these challenges, Explainable AI (XAI) and large language models (LLMs) have emerged as complementary approaches to bridge the gap between model complexity and human interpretability. While XAI focuses on attributing model decisions to input features, large language models introduce the possibility of generating natural language rationales. However, such explanations are not necessarily faithful to the underlying decision process and may reflect post-hoc plausible reasoning rather than true causal factors.

2.2. LLMS MODELS

While BERT relies on fully bidirectional self-attention, allowing each token to attend to all others within the input window [45], decoder-only Transformers introduce a causal masking constraint that restricts each token to attend only to preceding tokens. This archi-

tectural difference induces a distinct probabilistic factorization of the sequence and leads to autoregressive pretraining objectives [56].

$$P(x_1, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_{<t}) \tag{2.17}$$

In this formulation, the model is trained to predict the next token conditioned on its preceding context, which naturally enables open-ended text generation. Figure 2.6 illustrates the decoder-only Transformer architecture underlying modern large language models.

A major driver of performance improvements in these models is the empirical observation of scaling laws. These laws show that performance improves predictably with increases in compute, dataset size, and parameter count. This insight has triggered a global trend toward training increasingly large autoregressive language models [57].

Beyond model scale alone, post-training alignment techniques further improve LLM usability. Instruction tuning adapts pretrained models using curated instruction–response pairs, improving task-following behavior [58]. Reinforcement Learning from Human Feedback (RLHF) further refines model behavior by incorporating human preferences through reward models trained on human annotations.

Large Language Models represent a further evolution of Transformer-based architectures, characterized by a decoder-only design, massive scale, and extensive pretraining on large corpora. Unlike LSTM and BERT models, LLMs are typically deployed as frozen pretrained systems, with their behavior primarily influenced by alignment procedures and inference-time parameters rather than task-specific training. For consistency with previous model descriptions,

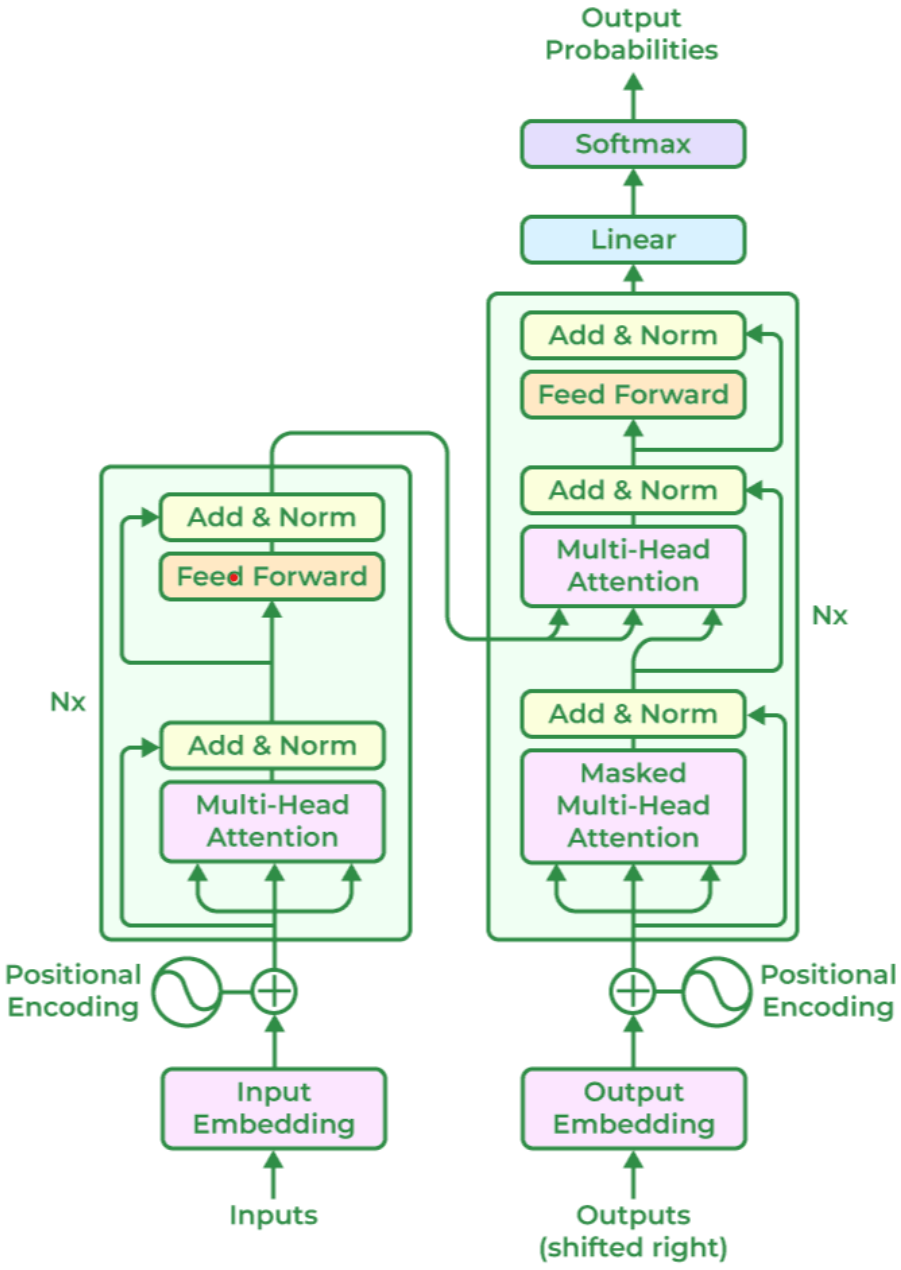


Figure 2.6 : LLM architecture

LLM parameters are organized into architectural, pretraining and alignment, and inference control parameters. Table 2.4 presents a detail view of large language models parameters.

2.2.1. ARCHITECTURAL PARAMETERS

- **Number of Transformer Layers (Depth)**: defines the depth of the LLM and directly impacts its reasoning capacity and representational hierarchy. Each layer consists of self-attention and feed-forward submodules that progressively refine token representations. Increasing depth enables the model to capture more abstract patterns and complex dependencies but substantially increases computational cost and memory requirements. Modern LLMs typically employ several dozen layers, reflecting a trade-off between expressive power and scalability.

- **Hidden Size**: specifies the dimensionality of internal token representations throughout the network. This parameter plays a central role in determining the model's capacity to encode semantic, syntactic, and contextual information. Larger hidden dimensions allow for richer representations and improved downstream performance but significantly increase the number of trainable parameters, leading to higher training and inference costs.

- **Number of Attention Heads**: controls how many parallel self-attention mechanisms operate within each Transformer layer. Each attention head attends to the input sequence from a different representational subspace, enabling the model to capture diverse contextual relationships such as long-range dependencies and semantic associations. While increasing the number of heads enhances modeling flexibility, it also increases computational complexity.

- **Context Window**: defines the maximum number of tokens that the model can process simultaneously. This parameter limits the range of information available for reasoning and long-range dependency modeling. Larger context windows enable improved coherence over long documents and multi-turn interactions but introduce significant memory and computational challenges due to the quadratic complexity of self-attention.

- **Vocabulary Size**: determines the granularity of tokenization used by the model. LLMs typically rely on subword-based tokenization schemes that balance lexical coverage and computational efficiency. Larger vocabularies reduce token fragmentation for rare words but increase embedding size and memory usage, whereas smaller vocabularies improve efficiency at the cost of longer token sequences.

2.2.2. PRETRAINING AND ALIGNMENT PARAMETERS

- **Pretraining Data Scale**: LLMs are pretrained on extremely large text corpora, often comprising hundreds of billions to trillions of tokens. The scale of pretraining data is a key factor driving generalization performance, linguistic competence, and factual coverage. Extensive pretraining enables LLMs to acquire broad world knowledge and emergent abilities without explicit supervision.

- **Pretraining Objective**: The primary pretraining objective for LLMs is autoregressive language modeling, where the model learns to predict the next token given all previous tokens in the sequence. This objective encourages the model to learn rich statistical representations of language structure and semantics, forming the foundation for downstream capabilities such as question answering and text generation.

- **Instruction Tuning**: Instruction tuning refers to a supervised fine-tuning stage in which the model is trained on datasets consisting of instruction–response pairs. This process

aligns the model's behavior with human expectations and improves its ability to follow natural language prompts. Instruction tuning plays a critical role in bridging the gap between general language modeling and practical task-oriented usage.

- **Fine-Tuning and Alignment Strategies**: Beyond supervised fine-tuning, alignment techniques such as Reinforcement Learning from Human Feedback (RLHF) are often employed to further refine model behavior. These strategies optimize the model with respect to human preference signals, improving safety, helpfulness, and response quality. Alignment procedures are essential for deploying LLMs in real-world interactive settings.

2.2.3. INFERENCE AND GENERATION PARAMETERS

- **Decoding Temperature**: controls the randomness of token sampling during text generation. Lower temperatures result in more deterministic and conservative outputs, while higher temperatures increase variability and creativity. Temperature selection directly affects output diversity and consistency, making it a crucial parameter for controlling model behavior at inference time.

- **Top-p (Nucleus Sampling)**: limits token selection to the smallest set of tokens whose cumulative probability exceeds a predefined threshold. This approach balances output diversity and coherence by filtering unlikely tokens while preserving flexibility in generation. Nucleus sampling is widely used in LLM inference to avoid degenerate or repetitive outputs.

- **Maximum Output Tokens**: specifies the upper bound on the generated sequence length. This parameter controls response verbosity and computational cost. Limiting output

length is particularly important in interactive or resource-constrained environments to ensure predictable response behavior.

- **Inference Interface**: LLMs are commonly accessed through abstracted inference interfaces, such as cloud-based APIs, which encapsulate decoding strategies and resource management. These interfaces provide controlled access to model parameters and generation settings, facilitating deployment while abstracting implementation details from end users.

The discovery of empirical scaling laws and the effectiveness of autoregressive pretraining ignited an international race to develop increasingly capable decoder-only large language models. Over the period from 2023 to 2026, several LLMs emerge such as closed or proprietary model such as ChatGPT, Claude and Gemini, and open-sources model such as LLaMA, DeepSeek and Mistral.

OpenAI's GPT series has undergone several significant leaps in capability, particularly with the transition from GPT-4o to the GPT-5 family in 2025–2026. By mid-decade, ChatGPT deployments relied on GPT-4o and GPT-5.x variants, each delivering notable improvements in reasoning, speed, and multimodal integration. GPT-5.2 expanded the model's context window to approximately 400,000 tokens and significantly reduced hallucination rates to around 4.8%, improving upon the 11–15% observed in earlier GPT-4 generation models. These releases also show state-of-the-art performance on benchmarks such as the AIME 2025 mathematics competition (achieving 100% accuracy) and the GPQA Diamond graduate-level reasoning suite [59].

Anthropic's Claude models are designed with a strong emphasis on safety-focused optimization and structured reasoning. The Claude 4.5 generation, released in late 2025 in Sonnet and Opus variants, represented a major advance in both capability and alignment [59].

Table 2.4 : Unified Large Language Model Parameters

Parameters	Role and Impact	Typical Values
Architecture		
Number of layers	Defines Transformer decoder depth and hierarchical reasoning capacity	24–80
Hidden size	Dimensionality of internal token representations controlling model capacity	2048–8192
Number of attention heads	Enables parallel self-attention over multiple representation subspaces	16–64
Context window	Maximum number of tokens processed simultaneously by the model	4k–128k+
Vocabulary size	Granularity of tokenization affecting lexical coverage and efficiency	32k–128k
Pretraining Data and Alignment		
Pretraining data scale	Amount of textual data used during large-scale self-supervised training	10^{11}–10^{13} tokens
Pretraining objective	Learning paradigm for next-token prediction and representation learning	Autoregressive language modeling
Instruction tuning	Alignment of the model to follow natural language instructions	Applied / Not applied
Fine-tuning strategy	Post-pretraining alignment and optimization for human interaction	SFT, RLHF
Inference and Generation		
Decoding temperature	Controls randomness and creativity in token generation	0.0–1.0
Top-p (nucleus sampling)	Limits candidate token set during probabilistic decoding	0.8–1.0
Maximum output tokens	Upper bound on generated response length	128–1024
Inference interface	Deployment and access abstraction for model interaction	Cloud API

A key feature of Claude 4.5 is the introduction of million-token context windows, provided through API beta modes, which enable long-document processing and complex multi-step reasoning workflows. On software engineering benchmarks, Claude 4.5 achieved a SWE-bench Verified score of 77.2%, outperforming several competing frontier models in code comprehension and automated bug fixing [60].

In parallel to the rapid evolution of proprietary LLMs, the period from 2023 to 2026 also witnessed the maturation of open-weight large language models, which played a central role in academic research due to their transparency, reproducibility, and accessibility.

Google's Gemini models, successors to LaMDA and PaLM 2, were developed as natively multimodal LLMs capable of jointly processing text, images, audio, video, and code. By 2025, Gemini 2.5 Pro and Gemini 2.5 Flash emerged as Google's flagship models, offering strong factual accuracy, high inference speed, and deep integration across the Google ecosystem (Docs, Gmail, Workspace) [61]. In November 2025, Google DeepMind released Gemini 3 Pro, with one-million-token context window, Mixture of Experts (MoE)-based scalability optimized for long-context reasoning, and state-of-the-art reasoning performance. Independent evaluations showed Gemini 3 excels at extremely long-context tasks and multimodal reasoning scenarios that earlier Claude and GPT systems struggled with [60].

Meta's LLaMA family marked a turning point in open-source LLM development. Starting with LLaMA 2 and progressing through LLaMA 3 and LLaMA 4, these models show that open-weight architectures could approach or match the performance of proprietary systems on a wide range of reasoning and language understanding tasks. LLaMA 3 emphasized improved instruction following and safety alignment, while LLaMA 4 further scaled model capacity and reasoning robustness, making the family a strong foundation for zero-shot and explanation-oriented evaluations.

Similarly, the Mistral models prioritized architectural efficiency and strong performance at moderate parameter scales. Mistral-7B and its successors showed that compact decoder-only models could deliver competitive reasoning and generation quality while remaining computationally efficient. This balance made Mistral models particularly attractive for experimental research settings, where controlled evaluation and repeated inference are required.

These open-weight models enable systematic investigation of large language model behavior under reproducible conditions, without the confounding effects of closed-system updates, proprietary fine-tuning, or undocumented alignment interventions. As such, they form the methodological basis for the large language model experiments presented in this dissertation.

Although BERT and LLMs originate from different architectural families, recent developments in deep learning are causing their trajectories to converge conceptually, methodologically, and functionally. Despite being encoder-only (BERT) or decoder-only (GPT-style models), both rely on: self-attention mechanisms, multi-head attention, positional encoding and large-scale pretraining on massive corpora. Thus, they share the same computational backbone, differing only in how attention masks are applied. Originally, BERT used MLM and GPT used autoregressive next-token prediction. But, decoder-only models can approximate bidirectional context by enlarging training windows and using deeper layers. Many recent encoder models incorporate autoregressive auxiliary losses. As a result, many tasks once dominated by BERT are now solved by LLMs, while LLMs increasingly benefit from encoder innovations. The discovery of empirical scaling laws shows that both encoder-based and decoder-based Transformers follow predictable power-law improvements as model size, dataset size, and compute increase. Scaling laws, documented in Kaplan et al. [57] in 2020, show that performance continues to improve smoothly across orders of magnitude, regardless of whether the architecture is encoder-only or decoder-only.

2.2.4. LLMS EVALUATION METRICS

The existing literature indicates that statistical evaluation frameworks originally developed for machine translation and text summarization have been successfully adapted to assess the quality of explanations generated by LLMs. These metrics provide quantitative estimates of lexical overlap, semantic similarity, and structural coherence between model-generated explanations and human-written reference texts. Although they were not specifically designed for fake news detection, such evaluation strategies offer valuable methodological insights that are directly applicable to XAI in this domain.

LLM evaluation metrics are used to assess model performance along various dimensions. As shown in Figure 2.7, they can be grouped into two main categories: statistical evaluation criteria and model-based evaluation criteria [62].

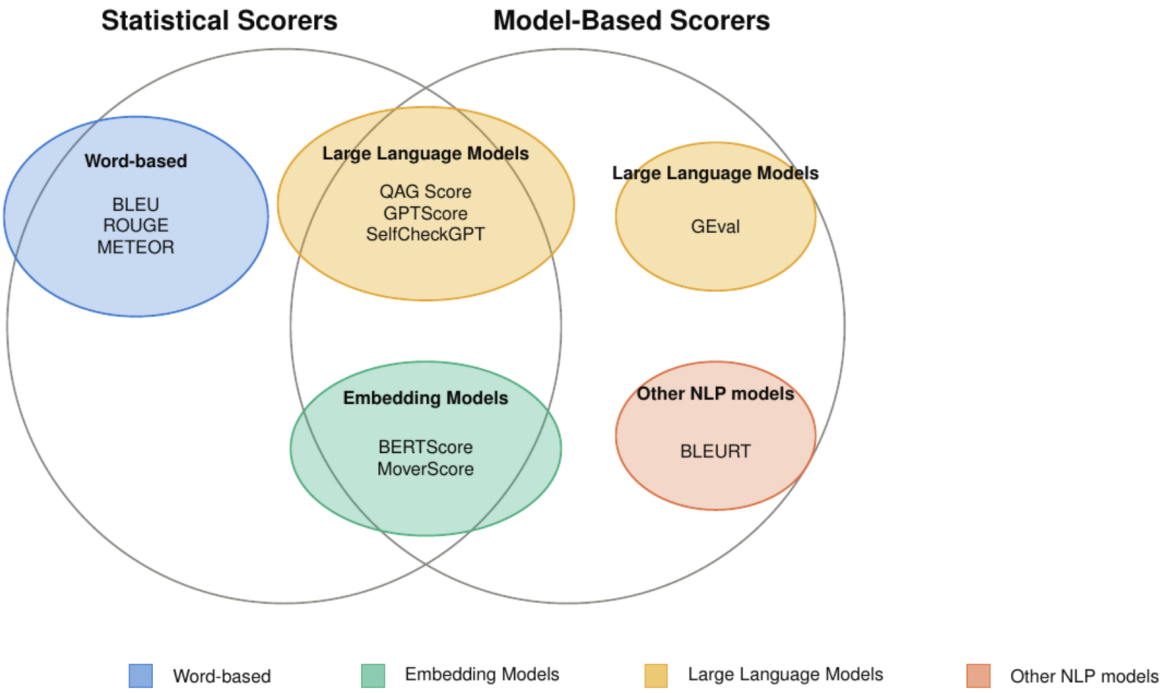


Figure 2.7 : LLM evaluation metrics

The existing literature shows that **statistical evaluation** frameworks originally developed for machine translation and summarization have been adapted to assess the quality of LLM-generated explanations. These metrics provide quantifiable measures of lexical overlap, semantic similarity, and structural coherence between model outputs and human-written reference explanations. These evaluation approaches offer valuable methodological insights applicable to our research.

The most commonly used statistical metrics include BLEU, METEOR, and ROUGE, each capturing different aspects of textual similarity.

BLEU (Bilingual Evaluation Understudy) (2002) measures the degree of lexical overlap between generated and reference texts by calculating the precision of matching n-grams. BLEU evaluates how many words and sequences of words in the generated explanation appear

in the reference explanation while applying a brevity penalty to discourage excessively short outputs. Higher BLEU scores indicate stronger lexical correspondence. Although BLEU remains widely used due to its simplicity and reproducibility, it tends to emphasize surface-level similarity and may underestimate semantically correct explanations expressed with different wording.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) (2005) was introduced to address some limitations of BLEU by incorporating both precision and recall during evaluation. Unlike BLEU, METEOR aligns generated and reference texts using exact word matching as well as stemming and synonym matching, making it more sensitive to semantic equivalence. The metric also includes a fragmentation penalty to account for differences in word ordering and fluency. As a result, METEOR often demonstrates stronger correlation with human judgments when assessing explanation quality.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (2004) was originally designed for automatic summarization evaluation and focuses primarily on recall, measuring how much information from the reference text is preserved in the generated output. Several variants exist, including ROUGE-N, which evaluates overlapping n-grams, and ROUGE-L, which measures the longest common subsequence between texts. In the context of LLM-generated explanations, ROUGE is particularly useful for assessing content coverage and determining whether key explanatory elements are retained relative to human-written references.

Since purely statistical scorers hardly not take any semantics into account and have extremely limited reasoning capabilities, they are not accurate enough for evaluating LLMs outputs that are often long and complex [63]. Moreover, most of these evaluation metrics assume the availability of a high-quality reference text, which is rarely the case in real-world settings. In practice, many LLM applications such as open-ended question answering, summa-

rization, or decision support do not admit a single ground-truth output, making reference-based evaluation infeasible or even conceptually ill-defined.

This limitation has prompted a growing body of work on reference-free and model-based evaluation methods, which aim to assess semantic correctness, reasoning quality, and usefulness without relying on a gold-standard output.

Model-based evaluation approaches rely on pretrained language models—often large language models themselves—to assess the quality of generated outputs. Unlike statistical metrics that compare outputs to reference texts, these methods evaluate semantic correctness, reasoning quality, coherence, and usefulness in a reference-free or weakly reference-dependent manner. They are increasingly used for tasks such as explanation evaluation, summarization quality assessment, and misinformation detection due to their flexibility and higher correlation with human judgments.

Beyond reference-based metrics, recent evaluation paradigms have increasingly leveraged LLM-based judges and factual consistency metrics. In these approaches, powerful LLMs are used to act as evaluators, scoring explanations according to criteria such as factual correctness, logical coherence, relevance to the prediction, and alignment with external evidence. This shift reflects a broader trend away from purely surface-level similarity metrics toward evaluation frameworks that emphasize semantic faithfulness and epistemic validity. Such metrics are particularly relevant where explanations must not only be linguistically plausible but also factually consistent with verifiable information.

LLM-based Judges. LLM-based judges, such as GPT-4, are prompted to evaluate generated explanations along dimensions including coherence, relevance, and factual correctness, and have been shown to correlate strongly with human judgments [64, 65]. Complementarily, factual consistency metrics such as FactCC [66], QAGS [67], and FEQA [68] evaluate whether

generated explanations remain faithful to the source content. These approaches are particularly relevant for explainable fake news detection, where explanations must be not only fluent but also factually grounded and epistemically sound.

QAG Score evaluates generated text by transforming it into a question-answering task. The method first generates a set of questions from either the reference text or the candidate output, and then uses a model to verify whether the relevant information is preserved in the generated response. This enables a more structured assessment of semantic adequacy and factual consistency. QAG is especially useful for evaluating explanation quality and factual grounding, as it focuses on whether key information can be correctly recovered. However, its effectiveness depends on the quality of the automatically generated questions and the robustness of the QA model used for scoring.

GPTScore is a reference-free evaluation framework that uses a pretrained language model to directly estimate the quality of generated text based on its probability of being a high-quality output. Instead of comparing a candidate output to a reference text, GPTScore conditions a language model on an evaluation prompt and computes a score based on token-level likelihoods. This approach leverages the internal knowledge and linguistic priors of large language models to approximate human judgment. GPTScore is particularly effective for evaluating fluency, coherence, and general text quality, but its performance is sensitive to prompt formulation and may reflect the biases of the underlying model.

G-Eval, introduced in 2023, is one of the first systematic frameworks to formalize the use of GPT-4 as an evaluation model for natural language generation tasks. It introduces structured prompts that guide the LLM to evaluate outputs across multiple quality dimensions, such as coherence, relevance, and correctness. The study of Liu et al. [64] shows that G-Eval achieves stronger correlation with human judgments compared to traditional automatic metrics.

This makes it particularly relevant for evaluating generated explanations, where assessing reasoning clarity and factual accuracy is more important than measuring lexical similarity. G-Eval establishes LLM-based judging as a viable and reliable evaluation paradigm.

- **Accuracy**. Answers contain verifiable facts. Sources are cited reliable. Dates and figures are correct.
- **Consistency**. Logic and coherence of reasoning. Presence of examples or evidence to support the explanation. Clear identification of misleading elements in the case of fake news.
- **Cohesion**. How clearly the explanation is presented and how easy it is to understand.
- **Relevance**. All important aspects are covered. The answer is detailed enough.

G-Eval introduces two methodological components:

- **Chain-of-thought evaluation**: The LLM first explains its reasoning before giving a score.
- **Form-based scoring**: The model fills out a predefined evaluation structure (e.g., criteria + final score).

Beyond specific implementations such as GPTScore and QAG, model-based evaluation methods share a common principle: they rely on pretrained language models to assess outputs without requiring explicit gold-standard references. These methods evaluate semantic correctness, informativeness, and reasoning quality through learned representations and probabilistic scoring mechanisms. Compared to statistical metrics, they better capture meaning-level similarity and are more aligned with human judgment, particularly in complex tasks such as

explanation generation and misinformation detection. However, they introduce new challenges such as model bias, prompt sensitivity, and limited interpretability.

Based on the latest 2026 benchmark analyses, GPT-5.2 offers the greatest stability, defined as consistency across reasoning tasks, reduced hallucination rates, and reliable performance across diverse evaluations. GPT-5.2 ranks first in reasoning performance in early 2026 benchmark suites and shows lower hallucination rates, 65% fewer hallucinations than GPT-4o on advanced reasoning tasks [69]. It is also rated one of the top overall LLMs in 2026 comparison rankings, with exceptionally high reasoning and factual-accuracy scores [70].

2.3. COGNITIVE AND AFFECTIVE PROCESSING

While computational architectures, ranging from recurrent networks like LSTMs to advanced Transformer-based models such as BERT and LLMs, show remarkable capabilities in automatically processing and classifying textual data, they fundamentally operate as abstract mathematical approximations of language. Evaluating their efficacy in real-world contexts, particularly in domains involving fake news, remains incomplete without examining how human users interact with and perceive such content. Consequently, shifting the focus toward human-centered perspectives is essential. By investigating human cognitive responses, information processing behaviors, and underlying decision-making mechanisms, researchers can establish a grounded baseline that connects automated algorithmic evaluations with the empirical realities of human cognition.

2.3.1. DUAL PROCESS THEORY

Humans remain the most critical yet vulnerable component in the information-processing chain, as the detection of fake news continues to pose significant cognitive challenges [20].

This vulnerability is consistent with established research on truth bias, which shows that individuals tend to initially accept information as true unless presented with strong and explicit contradictory evidence [21]. Consequently, skeptical evaluation does not occur spontaneously and typically requires deliberate cognitive effort and analytical engagement.

This phenomenon is commonly explained through **Dual Process Theory** [71], which proposes that human cognition operates through two interacting systems: an intuitive, fast, and automatic system (System 1), and a slower, deliberative, and analytical system (System 2).

System 1 corresponds to a fast, automatic, and heuristic-based mode of cognition. It operates with minimal conscious effort and enables rapid judgments based on associative learning, prior experience, and environmental cues. This system is highly efficient and allows individuals to process large amounts of information under time and cognitive resource constraints. However, due to its reliance on cognitive shortcuts, System 1 does not systematically engage in logical validation or evidential reasoning, instead favoring approximate judgments that prioritize speed over accuracy. As a result, it constitutes the primary mechanism through which initial credibility assessments of information are formed.

System 2, in contrast, is characterized by slow, effortful, and controlled processing. It is responsible for analytical reasoning, critical evaluation, and the deliberate verification of information. Unlike System 1, it requires the allocation of significant cognitive resources and conscious attention. System 2 is therefore typically engaged only when individuals detect inconsistencies, uncertainty, or when motivated to critically evaluate incoming information. In the context of information evaluation, System 2 plays a corrective role by potentially overriding intuitive judgments generated by System 1.

The interaction between these two systems gives rise to a fundamental asymmetry in human information processing. Due to an evolutionary preference for cognitive efficiency,

individuals tend to prioritize rapid acceptance of information over its critical rejection. Engaging in analytical verification requires sustained cognitive effort, which is often avoided in high-volume and time-constrained information environments. Furthermore, integrating new information that contradicts existing knowledge structures imposes additional cognitive cost, particularly when such information aligns with prior beliefs, thereby reducing the likelihood of critical reassessment.

This dual-system architecture provides a theoretical foundation for understanding systematic deviations from rational information processing and explains why individuals are particularly susceptible to heuristic-driven judgments in complex informational environments such as online news consumption.

2.3.2. IMPACT OF FAKE NEWS ON SYSTEM 1

The rapid and intuitive nature of System 1 processing makes it particularly vulnerable to manipulation within digital information environments. Fake news exploits heuristic-based cognition by presenting information in ways that maximize cognitive fluency, emotional salience, and rapid plausibility judgments while minimizing analytical scrutiny. Because System 1 prioritizes efficiency over verification, individuals frequently rely on simplified mental shortcuts when evaluating the credibility of online information.

One of the most influential mechanisms underlying this vulnerability is the use of **cognitive heuristics**. Heuristics are simplified cognitive strategies that enable individuals to make rapid judgments under conditions of uncertainty and limited cognitive resources. Although these mechanisms are generally adaptive in everyday decision-making, they may also produce systematic distortions in reasoning when individuals are exposed to misleading or manipulative information environments.

Among the most prominent heuristics involved in fake news reception is **confirmation bias**, whereby individuals preferentially attend to and accept information that is consistent with their pre-existing beliefs while disregarding contradictory evidence. This selective processing reinforces ideological consistency and reduces the likelihood of analytical reassessment. Similarly, the **bandwagon effect** increases the perceived credibility of information that appears socially validated or widely shared, particularly on social media platforms where popularity metrics strongly influence perceived trustworthiness.

Another important mechanism is the **anchoring bias**, in which early exposure to an initial claim disproportionately shapes subsequent interpretation and judgment. Even when corrective information is later introduced, the initial cognitive anchor often continues to influence reasoning processes. This phenomenon contributes significantly to the persistence of misinformation and the difficulty associated with correcting false beliefs once they have been cognitively encoded.

Fake news also heavily exploits the **familiarity bias** and the related **illusory truth effect**. Repeated exposure to the same information progressively increases its perceived plausibility, even when the information is objectively false. According to fluency-based theories of cognition, repeated or familiar content is processed more easily by the brain, creating a subjective sense of truthfulness through cognitive ease rather than factual verification [72]. This mechanism is particularly problematic in digital ecosystems characterized by algorithmic amplification and repetitive exposure to viral content.

A related phenomenon is the **linguistic familiarity heuristic**, whereby familiar semantic elements such as political figures, celebrities, organizations, or recognizable entities induce cognitive fluency and rapid plausibility judgments. The presence of well-known proper nouns may therefore reduce analytical scrutiny by creating an implicit perception of credibility or

coherence. In the context of fake news, these semantic anchors can bias attentional allocation and influence how individuals visually process information before conscious reasoning mechanisms are activated.

Beyond cognitive heuristics, fake news frequently leverages emotionally charged content to accelerate heuristic processing. High-arousal emotional stimuli such as fear, anger, outrage, or surprise can hijack attentional systems and reduce the probability of reflective evaluation. Emotional activation therefore reinforces System 1 dominance by encouraging rapid reactions and impulsive sharing behaviors rather than deliberate verification.

As illustrated in Figure 2.8, exposure to fake news typically initiates a rapid heuristic pathway dominated by System 1 processing. Under conditions of cognitive overload, information fatigue, or continuous exposure to high-volume online content, the activation of System 2 becomes increasingly difficult. Consequently, individuals are more likely to default to heuristic judgments, thereby increasing susceptibility to fake news.

2.3.3. COGNITIVE AND AFFECTIVE PROCESSING IN FAKE NEWS

While Dual Process Theory explains how individuals evaluate information through heuristic and analytical reasoning mechanisms, the processing of fake news cannot be fully understood without considering the role of affective processing. Human cognition encompasses a broad range of mental operations including perception, memory, interpretation, attention, and reasoning. Within this framework, information evaluation emerges from the continuous interaction between cognitive and affective systems [73].

Cognitive processing refers to deliberate and analytical mental operations involved in reasoning, fact-checking, logical verification, and evidence evaluation. These processes are

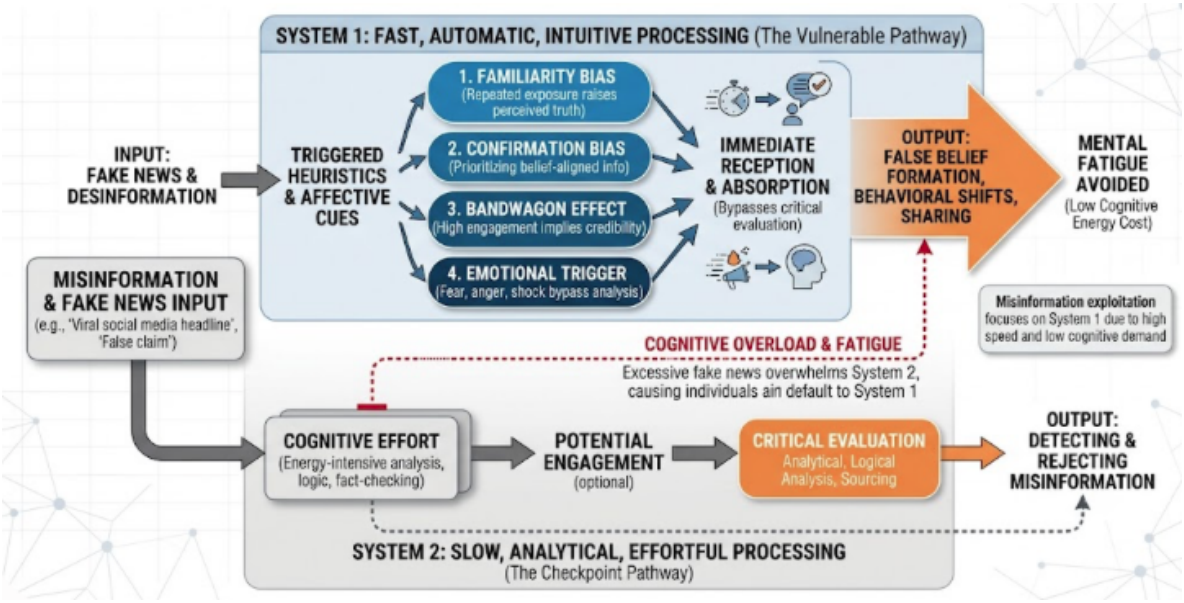


Figure 2.8 : Differential cognitive processing of fake news and the mechanisms of System 1 absorption.

primarily associated with System 2 activation and require sustained attentional resources and conscious cognitive engagement. In the context of fake news detection, cognitive processing enables individuals to critically assess the plausibility, internal consistency, and credibility of information rather than relying solely on intuitive impressions. Research conducted by Pennycook and Rand [74] suggests that susceptibility to fake news is largely associated with insufficient analytical engagement rather than purely ideological motivations, indicating that failures in reflective reasoning play a central role in misinformation acceptance.

In contrast, **affective processing** involves rapid, automatic, and emotionally driven evaluations associated with motivational relevance, emotional salience, and subjective impressions [75]. These affective responses often emerge prior to conscious reasoning and directly influence attentional allocation, memory encoding, and decision-making processes. Emotional

reactions therefore operate in close interaction with System 1 processing, shaping intuitive judgments before analytical verification mechanisms can intervene.

The relationship between emotion and fake news is particularly significant in digital environments where emotionally charged content is algorithmically amplified to maximize engagement. Vosoughi et al. [76] show that false information spreads significantly faster and reaches larger audiences than truthful information on social media platforms, largely because fake news evokes stronger emotional reactions such as surprise, fear, disgust, and outrage. Similarly, Osmundsen et al. [77] showed that high-arousal emotions, particularly anger, substantially increase user engagement behaviors including sharing, commenting, and diffusion of political misinformation.

The interaction between cognitive and affective processing can be further understood through the **Appraisal Theory of Emotion** [78]. According to this framework, emotional responses arise from rapid cognitive evaluations of external stimuli in relation to an individual's goals, expectations, and perceived threats.

Importantly, these appraisals frequently occur outside conscious awareness and prior to deliberate reasoning. In the context of fake news exposure, emotionally salient or expectation-violating content may therefore trigger immediate affective reactions before individuals consciously analyze the validity of the information.

This interaction creates a critical vulnerability within online information ecosystems. Although affective responses may occasionally serve as intuitive warning signals when information appears incoherent or suspicious, highly emotional content can also overwhelm analytical reasoning processes and reinforce heuristic judgments. Under conditions of cognitive overload, emotional arousal, or continuous information exposure, System 1 processing

tends to dominate, thereby reducing the probability of reflective evaluation and increasing susceptibility to misinformation.

The cognitive and affective mechanisms involved in fake news reception are particularly important for understanding observable physiological responses during information evaluation. Emotional arousal, attentional engagement, cognitive conflict, and analytical effort can manifest through measurable behavioral and neurophysiological signals such as gaze allocation, fixation duration, pupil dilation, and variations in neural activity. Consequently, integrating cognitive and affective theories provides a theoretical foundation for the use of eye-tracking and EEG technologies to investigate how individuals process fake news in real time.

2.3.4. COGNITIVE EYE-TRACKING ANALYSIS

Given the central role of cognitive effort and affective processing in fake news evaluation, researchers have increasingly turned to physiological sensors to infer internal cognitive states. Visual attention constitutes a fundamental mechanism through which cognitive resources are selectively allocated during information processing.

Studies using multimodal physiological data such as galvanic skin response, Electroencephalography (EEG), and heart-rate variability have shown the feasibility of detecting emotional and cognitive reactions during information processing [79, 80]. Recent advances in affective computing highlight the value of eye-tracking technology for inferring emotional and cognitive states during complex tasks [81]. Eye-tracking systems capture fixation points, saccadic movements, gaze duration, and pupil diameter, all of which provide rich indicators of attention allocation and cognitive engagement.

Eye movements offer a direct window into the mechanisms of visual perception and attention. Fixations indicate the allocation of cognitive resources to specific elements of a

stimulus, while saccades reflect rapid shifts in attention. Pupil diameter, in particular, has been shown to correlate with cognitive load: Kahneman and Beatty [82] show that pupil dilation increases with task difficulty, reflecting the mental effort required to maintain information in working memory. In the context of misinformation, these physiological markers can reveal whether individuals experience cognitive conflict, heightened emotional arousal, or increased processing difficulty when confronted with deceptive content.

For these reasons, eye-tracking constitutes a powerful methodological tool for studying the cognitive and affective processes involved in fake news detection. When combined with EEG or other physiological measures, eye-tracking enables a multimodal analysis of attention, emotional reactivity, and cognitive load. Such an approach provides deeper insights into the mechanisms underlying susceptibility to fake news and supports the development of explainable AI systems capable of integrating human cognitive signals into fake news detection pipelines.

2.4. BACKGROUND SUMMARY

This chapter has established the foundational theoretical and methodological frameworks required to address the detection and cognitive impact of fake news. We explored the evolution of computational linguistics, shifting from the recurrent temporal modeling of LSTMs to the deep, bidirectional contextualization of BERT and advanced LLMs. These architectures provide the mathematical and algorithmic rigor necessary to parse the subtle, complex semantic structures often deployed in deceptive online narratives.

However, as emphasized in the latter half of this background, algorithmic performance alone cannot fully capture the reality of fake news dissemination. To bridge the gap between computational systems and human vulnerability, it is equally vital to understand the cognitive

mechanisms, human decision-making processes, and behavioral responses that dictate how individuals consume and evaluate text.

With these dual pillars computational models and human-centered cognitive perspectives firmly established, the document now turns to a critical review of the literature. The next chapter will examine how previous studies have operationalized these theories, highlighting the methodologies current researchers employ to detect fake news and identify the remaining gaps that this doctoral research aims to fulfill.

CHAPTER III

RELATED WORK

This chapter provides a critical and structured review of fake news detection research, with a focus on three complementary dimensions: (i) model performance, (ii) explainability, and (iii) human cognitive validation. Rather than purely listing prior work, this review aims to identify methodological trends, limitations, and open research gaps that motivate the contributions of this thesis.

3.1. FAKE NEWS DETECTION AND EXPLANATION STRATEGIES

While this chapter primarily focuses on fake news detection at the content and article levels, it also includes selected studies on influence operations and coordinated disinformation campaigns. These works do not directly address claim-level veracity classification but provide complementary insights into how fake news is produced, amplified, and coordinated at scale. Such approaches are particularly relevant for understanding campaign-level dynamics and motivate the use of clustering and pattern mining techniques, as explored later in this thesis.

3.1.1. EARLY APPROACHES AND DATASETS

Three early studies laid the foundations of fake news detection research by addressing complementary aspects of the problem. Shu et al. [20] (2017) introduced a unifying conceptual framework combining content, social context, and user behavior, while Wang [83] (2017) enabled systematic evaluation through the LIAR benchmark dataset. Vosoughi et al. [6] (2018) further show that fake news exhibits distinct diffusion dynamics driven primarily by human users.

Shu et al. [20] (2017) provided one of the earliest comprehensive overviews of fake news detection on social media from a data mining perspective. The authors analyzed the characteristics of fake news and proposed a conceptual framework that incorporates three main components: the content of the news, the social context in which it spreads, and the users who interact with it. Their work highlights the importance of combining textual analysis with social network information to effectively detect fake news. This survey laid the conceptual foundations for many subsequent studies in fake news detection. However, they didn't propose a specific detection model nor report empirical performance metrics. As a survey paper, their work reviews existing fake news detection approaches based on traditional machine learning classifiers such as SVM, logistic regression, and decision trees, primarily relying on handcrafted textual and social features. While the framework provides a foundational taxonomy of fake news detection strategies, it predates the adoption of deep neural networks and does not address representation learning, explainability, or quantitative evaluation of model performance.

Wang [83] (2017) introduced the LIAR dataset, one of the first publicly available benchmark datasets for fake news detection. The dataset contains more than 12,000 short political statements collected from the PolitiFact fact-checking website and labeled according to six levels of truthfulness. In addition to the statements themselves, the dataset includes metadata such as the speaker, political affiliation, and context of the claim. The LIAR dataset has since become a widely used benchmark for evaluating machine learning models designed to detect fake news. To establish baseline performance, the authors evaluated several machine learning models, including logistic regression, Support Vector Machines (SVM), CNNs, LSTMs, and BiLSTMs. The best-performing model, a BiLSTM, achieved an accuracy of approximately 27.4% on the six-class classification task, highlighting the intrinsic difficulty of fine-grained truthfulness prediction based on short political statements. These relatively

low performance scores suggest that short textual claims alone provide insufficient contextual and semantic information, motivating the need for richer representations, external context integration, and more advanced language models.

Vosoughi et al. [6] (2018) conducted a large-scale empirical study analyzing the diffusion patterns of true and false information on social media. Their dataset comprised approximately 126,000 news stories shared by nearly 3 million Twitter users between 2006 and 2017. The results show that false news spreads faster, deeper, and more broadly than true news, particularly in political contexts. Importantly, the study showed that human users, rather than automated bots, were primarily responsible for amplifying the spread of fake news, providing strong empirical evidence of distinct diffusion behaviors. However, the study does not propose a fake news detection model nor report classification performance metrics, as its primary objective is to characterize diffusion behavior rather than perform automated detection. While these findings provide valuable insights into human-driven disinformation dynamics, they remain descriptive and are not directly operationalized into detection models, limiting their applicability for real-time or early-stage fake news detection.

3.1.2. TRADITIONAL AND PROPAGATION-BASED MODELS

Shu et al. [7] (2019) proposed a tri-relationship model that jointly captures the interactions between publishers, news articles, and users. The TriFN model relies on multi-embedding representations derived from content features, user profiles, and interaction patterns. Evaluated on the FakeNewsNet dataset, including the PolitiFact subset containing 120 articles, the approach achieved an accuracy of 84.3% and approximately 80% detection performance within 48 hours, highlighting the importance of modeling social context. However, the model relies on handcrafted textual representations and pre-transformer feature engineering, and its evaluation is conducted on a relatively small dataset, which raises concerns regarding

scalability, generalization, and applicability to modern large-scale disinformation scenarios. Moreover, while TriFN effectively integrates social context, it does not address explainability nor validate whether the learned representations align with human cognitive processes.

Monti et al. [84] (2019) introduced a geometric deep learning framework for fake news detection on social media. Their approach models information propagation cascades using Graph Convolutional Networks (GCNs) applied to Twitter interaction graphs. Experiments conducted on a dataset comprising 5,976 Twitter news cascades achieved a Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) score of 92.7%, showing that propagation structures provide strong signals for early fake news detection, even within the first two hours after publication. However, the reliance on propagation structures limits applicability in scenarios where diffusion data is sparse, delayed, or unavailable, such as at the moment of publication. Moreover, the model does not provide explainability at the semantic level, nor does it assess whether the learned structural patterns align with human cognitive cues used during fake news evaluation. These results confirm that propagation-based models provide strong early signals for fake news detection. However, their dependence on social diffusion motivates complementary approaches that rely on content-based representations, particularly transformer-based models capable of operating at publication time.

Shu et al. [85] (2020) investigated hierarchical propagation networks for fake news detection by modeling the temporal and structural diffusion patterns of information on social media. The model uses structural features of propagation graphs, including node relationships and temporal dynamics. Experiments were conducted on the FakeNewsNet dataset, and the proposed Linguistic Inquiry and Word Count–Hybrid Psychological Feature Network (LIWC-HPFN) model achieved 87% accuracy, confirming the importance of temporal propagation patterns in fake news detection. However, the reliance on handcrafted psychological features and propagation data limits the model's scalability and applicability at early publication

stages, and the approach does not provide explicit semantic or human-centered explanations aligned with readers' cognitive processes. These findings further show the effectiveness of propagation-aware models. However, their dependence on diffusion data and predefined linguistic features motivates the exploration of transformer-based architectures capable of learning semantic representations directly from text at publication time.

Zhou et al. [86] (2020) proposed a theory-driven framework for early fake news detection that integrates writing style features derived from deception and persuasion theories. The model incorporates lexical, syntactic, and semantic characteristics, as well as clickbait-related attributes, to capture stylistic cues indicative of misinformation. The approach was evaluated on the PolitiFact (120 articles) and BuzzFeed (90 articles) datasets using traditional machine learning classifiers, including SVM, XGBoost, and Random Forest. The results show that stylistic features can achieve competitive detection performance in early-stage scenarios, confirming their usefulness when social propagation signals are not yet available. However, the evaluation is conducted on relatively small datasets, which limits the generalizability of the results. Moreover, the reliance on handcrafted stylistic features constrains the model's ability to capture deeper semantic representations and does not provide explicit explainability aligned with human cognitive processing. While theory-driven stylistic features are effective for early detection, their dependence on manual feature engineering motivates the adoption of deep neural models capable of learning semantic representations directly from raw text.

Another study from Shu et al. [87] (2021) proposed an early detection model capable of identifying fake news within five minutes of publication and before reaching 50 shares, achieving an accuracy of 90%. The model relies on the profile of the user and the analysis of the comments' content. They opted for the CNN algorithm and incorporated two new mechanisms to facilitate rapid detection. The training dataset, adapted from Twitter and Weibo, contained 4,600 articles, of which only 10% were labeled as fake. However, this approach

relies heavily on user profile characteristics, which may change over time, limiting the model's robustness since fake news spreaders are not always consistent across events. Furthermore, the dependency on user interactions makes the approach less applicable at the moment of publication and does not provide interpretable explanations grounded in the semantic content of the news itself.

3.1.3. DETECTION WITH BOTS AND TROLLS IDENTIFICATION

This line of research focuses on the identification of influence campaigns, particularly through the detection of trolls and bots. Trolls can be defined as humans who aim to generate controversy, unlike bots, which are automated agents designed to emulate human user behavior. Consequently, the studies reviewed below address influence operation detection at the account and campaign levels rather than fake news classification at the content level. As such, their performance relies on large-scale behavioral coordination signals that are not directly applicable to early-stage fake news detection or to explainable, content-centric analysis. Moreover, these approaches do not provide semantic or human-interpretable explanations at the textual level, nor do they assess whether the detected patterns align with the cognitive cues used by readers when evaluating news credibility.

Alizadeh et al. [88] (2020) analyzed content-based features to predict social media influence operations, focusing on troll accounts from different geopolitical contexts. The dataset contained 2,660 Chinese troll accounts generating over 1.9 million posts. The authors used Random Forest classifiers on post–URL pair features and temporal activity patterns, showing that coordinated messaging and link-sharing behavior can reveal organized fake news campaigns. The model achieved high performance, with strong precision and F1-scores in detecting coordinated behaviors among troll accounts. These results show that behavioral and URL-sharing patterns are effective signals for identifying influence operations.

Alhazbi [89] (2020) introduced a behavior-based machine learning approach to identify state-sponsored troll accounts on Twitter. The dataset included 1,681 Saudi troll accounts, and the model relied on behavioral indicators such as number of tweets, retweets, hashtags, and URLs. Using supervised learning classifiers, the method achieved 94.4% accuracy for Saudi troll detection and 72.5% for Russian trolls, highlighting the importance of behavioral patterns in identifying coordinated influence operations.

Heidari et al. [90] (2020) conducted an empirical study on machine learning algorithms for social media bot detection. Using the Cresci 2017 dataset with over 5,500 bots and 3,300 fake accounts, the authors evaluated multiple models including Random Forest, Artificial Neural Networks (ANN), and Support Vector Machines (SVM). The analysis incorporated sentiment-based textual features, showing that sentiment and linguistic signals can improve classification performance for identifying bot-generated fake news.

Smith et al. [91] (2021) proposed a machine learning approach to identify influential actors in disinformation networks on Twitter. Using the Cresci 2017 dataset, which contains 3,151 bot accounts, the study analyzes behavioral, linguistic, and topic-based features extracted from user activity. A Random Forest classifier achieved a precision of 96%, showing the effectiveness of combining behavioral and linguistic signals to identify influential bots.

Overall, these studies show the effectiveness for identifying coordinated disinformation actors. However, their reliance on user-level signals limits their applicability for early detection and does not address the interpretability of the disinformation itself.

3.1.4. TRANSFORMER-BASED MODELS FOR FAKE NEWS DETECTION

After the emergence of traditional machine learning approaches, the introduction of BERT in 2019 marked a major turning point in fake news detection research. Transformer-based

encoders such as BERT show superior performance by capturing deep semantic and contextual information from text. This subsection reviews studies that employ BERT primarily for content-based fake news detection. We also observe that several works extend BERT by integrating additional algorithms to incorporate contextual information such as user behavior and dissemination patterns.

Jwa et al. [92] (2019) introduced exBAKE, an automatic fake news detection model that leverages the bidirectional representation capabilities of BERT to capture deep contextual relationships within news articles. Unlike traditional models that often rely on shallow linguistic features, exBAKE focuses on the semantic coherence between headlines and body text to identify inconsistencies typical of clickbait and disinformation. By fine-tuning the pretrained BERT architecture on large-scale datasets, the authors show that their approach significantly outperforms previous state-of-the-art methods, such as CNN and LSTM, with a F1-score of 74.6%. However, despite the use of contextual embeddings, the model focuses exclusively on textual coherence and does not incorporate social or propagation context, which are known to be critical for fake news detection in real-world scenarios. Furthermore, the approach does not provide interpretable explanations, limiting its applicability in settings where transparency and human understanding are required.

Heidari et al. [93] (2021) proposed a BERT-based model for fake news detection that incorporates social bot activity features during the COVID-19 pandemic (2350 tweets). The model uses BERT for textual representation of tweets and integrated "botometer" to obtain a probability score of being a bot. The core innovation of the framework proposed by the authors lies in its hybrid architecture, which supplements the semantic depth of BERT with structural behavioral analysis. The proposed approach improves detection performance, achieving around 92% accuracy and an F1-score of approximately 91%, outperforming traditional machine learning classifiers that rely only on textual features. However, the reliance on

external bot detection tools and user-level signals limits the model's applicability in scenarios where such metadata is unavailable or unreliable. In addition, the approach depends on post-publication interaction data, reducing its effectiveness for real-time detection at the moment of publication. Furthermore, the model does not provide explicit interpretability mechanisms, leaving the decision-making process opaque despite its improved performance.

Kaliyar et al. [94] (2021) introduced FakeBERT, a deep-learning architecture for detecting fake news circulated during the 2016 U.S. Presidential Election, using a 20,800 item Kaggle dataset containing ID, title, author, and full text. The authors explore multiple input modalities, integrating BERT-base with convolutional neural networks (CNN) and LSTM components to capture both contextualized linguistic representations and local textual patterns. Experimental results showed strong performance across all tested architectures, with BERT+LSTM achieving 97.55% accuracy, while the full FakeBERT model reached 98.9% accuracy, establishing one of the strongest results reported in this benchmark. These results show the effectiveness of combining transformer-based contextual embeddings with sequential and convolutional neural networks to enhance textual representation learning. However, the exceptionally high performance reported may be influenced by the characteristics of the dataset, which is known to be relatively clean and less representative of real-world fake news scenarios. Moreover, the model focuses exclusively on textual features and does not incorporate social or contextual signals, which limits its robustness in dynamic environments. Finally, despite its high accuracy, the approach lacks interpretability, providing no insight into the reasoning behind its predictions.

Dou et al. [13] (2021) proposed User Preference-aware Fake News Detection (UPFD), a graph-based framework that integrates user preference signals with news content and propagation patterns. The model combines BERT for text encoding and GraphSAGE-based Graph Neural Networks to capture relationships between users and news. Experiments conducted

on the FakeNewsNet datasets (Politifact and GossipCop) show that UPFD achieves 84.62% accuracy and 84.65% F1 on Politifact, and 97.23% accuracy and 97.22% F1 on GossipCop, outperforming previous graph-based baselines. These results highlight the effectiveness of combining semantic text representations with user preference modeling and social context. However, the model relies on user interaction data and social network structures that may not be available at early stages of news dissemination, limiting its applicability for real-time detection. Furthermore, the integration of multiple components increases model complexity and reduces interpretability, making it difficult to understand which features drive the final decision. The approach also does not explicitly address alignment with human cognitive processes.

Kula et al. [95] (2021) evaluated several BERT-derived architectures to address disinformation using the widely referenced ISOT dataset, composed of 44,898 news items (21,417 real and 23,481 fake). Their dataset includes title, text, and publication date, enabling both textual and limited contextual analysis. The study compares models built on BERT, RoBERTa, and RNN-based architectures, examining how transformer-based contextual embeddings interact with recurrent neural layers. The authors report an average accuracy of 98.7%, confirming that transformer models, especially when combined with sequential neural architectures, yield strong and consistent performance in large-scale fake news detection. However, the exceptionally high performance reported on the ISOT dataset may be influenced by its relatively clean and well-separated structure, which does not fully reflect the complexity and ambiguity of real-world misinformation. Moreover, the study focuses primarily on textual features and does not incorporate social or propagation context. Despite the strong results, the models lack interpretability and do not provide insights into the underlying reasoning process, limiting their reliability in high-stakes decision-making scenarios.

Glazkova [96] (2021) focuses on fake news detection during the COVID-19 pandemic using the FakeCovid dataset, a multilingual, cross-domain collection of fact-checked COVID-19 posts. The dataset includes post ID and message, meaning the classification task relies heavily on textual content rather than metadata. The study experiments with BERT, RoBERTa, and the pandemic-specialized BERT-Twitter COVID model. The results show an average accuracy of 98.7%, showing that COVID-domain-adapted transformer models, when fine-tuning with a huge dataset, provide highly effective detection of pandemic-related fake news. However, the high performance may be influenced by the domain-specific nature of the dataset, which focuses exclusively on COVID-19-related fake news and may not generalize well to broader or more diverse fake news scenarios. Additionally, the reliance on textual features alone ignores social and propagation context, and the models do not provide interpretable explanations, limiting their applicability in real-world and human-centered settings.

Keya et al. [97] (2022) investigate fake news detection in the Bengali language by proposing the AugFake-BERT framework, which explicitly addresses the issue of class imbalance in fake news datasets. The study considers a highly skewed dataset containing 48,678 real news articles and only 1,299 fake ones, thereby motivating the use of data augmentation techniques. The dataset comprises news identifiers, titles, and textual content. The authors explore text-based deep learning architectures that combine BERT-base with CNN and LSTM models to enhance feature representation. By augmenting the minority class and leveraging contextual embeddings provided by BERT, the proposed approach achieves a final accuracy of 92.45%, showing the effectiveness of data augmentation strategies in low-resource and highly imbalanced settings. However, the performance gains are strongly tied to the effectiveness of the data augmentation process, which may introduce synthetic bias and affect the authenticity of the generated samples. Additionally, the model remains focused on textual features and does not incorporate social or contextual information. As with many

transformer-based approaches, it lacks interpretability and does not assess whether the learned representations align with human cognitive processes.

In summary, **transformer-based models have shown superior ability to capture semantic nuances compared to traditional RNN-based architectures** (see Table 3.1). Studies published between 2021 and 2022 indicate that the highest detection accuracies, often exceeding 98%, are achieved through structural hybridization, where BERT is combined with CNN or LSTM layers to enhance textual modeling. However, it is important to note that most of these works primarily rely on textual features. The majority of context-aware approaches integrate BERT with additional algorithms rather than exploiting BERT alone for contextual modeling. In this context, my 2024 contribution on BERT-based explainable fake news detection simultaneously integrates textual and contextual features within a unified architecture, while also incorporating explainability mechanisms.

Table 3.1 : BERT Studies - Fake news detection

Title	AugFake-BERT: handling imbalance through augmentation of fake news using BERT to enhance the performance of fake news classification *(Keya, 2022)*	FakeBERT: Fake news detection in social media with a BERTbased deep learning approach (*Kaliyar et al., 2021)*	Implementation of the BERTderived architectures to tackle disinformation challenges (*Kula et al.,* 2021*)*	User preference aware fake news detection (*Dou et al., 2021*)	BERT model for fake news detection based on social BOT activities in the COVID19 pandemic (*Heidari et al*., 2021)	Exploiting CT-BERT and ensembling learning for COVID19 fake news detection (*Glazkova, 2021*)	exBAKE: Automatic Fake News Detection ModelBased on BERT *(Jwa et al., 2019)*
Approach	Content	Content	Content	Content Context	Content Context	Content	Content
Dataset	Langue Bengali (Inde) 4000-4000 true and false story in Bangal language	Kaggle 20,800	ISOT (44,898 items 21,417 of which are real items and 23481 are fake items.	FakeNewsNet	COVID-19 2350 tweets	FakeCovid - A Multilingual Cross-domain Fact Check News Dataset for COVID-19	90,000 documents form CNN news and 197,000 doc from Daily Mail)
Features		ID, title, author, text and lable (fake or real)	Headlines and body texts (title and text) article title, text, article publication date	Historical posts User comments retweet count, friend count.		Post ID, the post, and label which is ”fake” or ”real	Headlines and body texts
Algorithm method	BERT base and CNN, LSTM	BERT base and CNN LSTM	BERT and RNN RoBERTa et RNN	FUSION: (1) BERT- pretrain historical tweets and GCN (graph convolution network)	BERT- bot or not bot helps to find fake news	Bert RoBERTa BERT-TwitterCOVID	BERT Linear fonction-Softmax for classify
Accuracy	AugFake: 92%	LSTM 97.55% FakeBERT 98.9%	Average of 98.7%	Up to 97%	92%	Average of 98.7%	75%

3.1.5. EXPLAINABLE FAKE NEWS

Several studies have explored the use of XAI methods to add explainability in machine learning-based fake news detection systems. However, the internal representations learned by BERT are distributed across millions of parameters, making the reasoning process inherently opaque. According to the **NIST 8312 framework**, *explanation* refers to the information provided by the system about how a decision was reached, whereas *interpretation* concerns the human ability to understand, contextualize, and derive meaning from that information. This distinction is essential: an explanation can be technically correct yet remain uninterpretable for analysts, auditors, or end-users.

This limitation highlights the necessity of integrating XAI extensions that not only generate faithful explanations but also support human interpretation by presenting insights in a form that aligns with cognitive expectations and domain expertise without compromising detection performance.

In response to this challenge, prior research has explored different model architectures and explanation strategies aimed at improving both the *explainability* of the underlying decision process and the *interpretability* of the resulting outputs for humans.

Shu et al. [98] (2019) proposed DeFEND model "Explainable Fake News Detection". This model scrutinizes user comments and the news text. User comments, which include text, hashtags, and emojis, are significant as they help to contextualize the comments. The model employs a RNN-based word encoder to learn sentence representations. For news sentence explainability, the authors utilized the hierarchical attention network (HAN), and for user comment explainability, they used HPA-BLSTM as baselines. While DeFEND offers explainability, it doesn't provide a probability score for the likelihood of fake news. Additionally, the study doesn't address the credibility of user comments themselves, which could be a

limitation. While dEFEND represents a major step toward explainable fake-news detection by jointly leveraging article content and user comments, its reliance on user-generated reactions introduces significant limitations. The availability, quality, and reliability of comments vary widely across platforms, making the model less robust and less generalizable in contexts where social feedback is sparse, noisy, or strategically manipulated.

Mohseni et al. [99] (2021) trained a BiLSTM network using Word2Vec word embeddings, relying on content-based features such as the news headline, article text, and source information. Their model achieved an accuracy of 73.65%. To enhance interpretability, the authors employed HAN to highlight sentence-level contributions and integrated several visual explanation tools: a heatmap for word-level feature attribution in headlines, an attribution score indicating the model's confidence in the veracity classification, bar charts comparing the contribution of different article segments, and a donut chart illustrating source-level attribution relative to both headline and content. The bar charts, for instance, displayed lower scores for article sections that were less relevant to the headline or less influential in the model's decision. The primary limitation of this study lies in the restricted size of the dataset, which also contains politically oriented content that may introduce bias. Furthermore, the analysis is limited to textual features only, overlooking contextual, social, and multimodal signals that are increasingly important in misinformation detection.

Szczepariski et al. [14] (2021) used a content-based approach with BERT, training their model on a dataset of 4,652 items and reporting an accuracy of 98%. They incorporated two XAI methods, LIME and Anchors, to interpret the model's predictions. In their LIME configuration, the explainability module highlighted the five most influential words contributing to the classification of a news item as true or false, along with their associated contribution weights. For example, in one of their explanations, the model predicted a 97% probability that the news was false, with words such as "gave" contributing 28% and "Hilary" contributing 21% to this

outcome. The main limitation of this study lies in the questionable relevance of the extracted keywords: the highlighted terms do not appear intrinsically meaningful for determining the falseness of the news. This suggests that keyword-based explanations may be insufficient for capturing the deeper mechanisms of disinformation, especially given that the contextual and multimodal dimensions of fake news remain largely underexplored.

Chien et al. [100] (2022) proposed XFlag, a system using LSTMs with Layer-wise Relevance Propagation (LRP) to explain fake news. Their work extends LRP to recurrent architectures, enabling the model to compute which input features, content, user information, and sentiment, are most relevant to the classification. While XFlag shows the potential of LRP for generating fine-grained explanations, its focus remains largely limited to textual content and user-level attributes, similar to the approach taken by Szczepański et al. [14].

A critical cross-analysis of existing explainable fake news detection models reveals a significant methodological gap: the lack of quantitative metrics to evaluate explanation quality. While models such as DeFEND [98] and XFlag [100] provide visual or hierarchical insights into model decision-making, their evaluations remain largely qualitative. For instance, the application of LIME by Szczepański et al. [14] highlights words with high attribution scores that lack inherent semantic links to fake news, raising concerns about the true interpretability of the features. Furthermore, without the integration of metrics like METEOR or empirical validation through eye-tracking, it remains impossible to determine if these 'explanations' truly align with human cognitive processes or merely represent mathematical artifacts of the neural network. This limitation motivates the exploration of LLM-based reasoning and human-centered evaluation frameworks, which aim to bridge the gap between model performance and human interpretability.

3.2. LLMS FOR FAKE NEWS

After 2021–2022, research on fake news detection began to shift toward improving the efficiency, scalability, and generalization capabilities of Transformer-based models. In particular, increasing attention has been directed toward LLMs, which have shown strong reasoning abilities and competitive performance across a wide range of natural language understanding tasks. Although LLMs are fundamentally based on the transformer architecture, similarly to BERT-type encoders, their substantially larger scale, emergent reasoning capabilities, and distinct training and inference paradigms motivate a separate line of investigation. For this reason, this chapter specifically reviews literature on fake news detection performance and explanation quality that leverage LLMs.

Chen & Shu [19] (2023) highlighted how LLMs, while offering potential tools to detect and counter disinformation, also pose new risks because of their ability to generate highly convincing misleading content. The authors discussed potential strategies to mitigate these risks, highlighting the need for a multidisciplinary approach. This study emphasized the need of metrics to represent the fidelity of an explanation, not just its persuasiveness. They adopted LLMs such as GPT-4 with zero-shot prompting strategies as the representative misinformation detectors to assess and compare the detection hardness of LLM generated misinformation and human-written misinformation. The experiment result analysis are two fold: First, they can observe that it is also generally hard for LLM detectors to detect ChatGPT-generated disinformation, especially those generated via Hallucinated News Generation, Totally Arbitrary Generation and Open-ended Generation. For example, LLM detectors can hardly detect fine-grained hallucinations.

Wei et al. [101] (2023) show that the reasoning quality of LLMs is significantly enhanced through Chain-of-Thought (CoT) prompting. By explicitly instructing the model to 'think

step-by-step,' this technique forces the LLM to decompose complex fact-checking tasks into a series of intermediate logical deductions before arriving at a final veracity label. This shift from direct answering to sequential reasoning provides a more granular basis for evaluating the faithfulness of an explanation, as the final justification is no longer a post-hoc rationalization but a direct reflection of the model's computational trajectory.

Krishna et al. [102] (2023) investigated the role of post-hoc explainability in LLMs, moving beyond the traditional use of explanations as purely interpretative tools. They proposed the AMPLIFY framework, which leverages post-hoc explanation methods to generate structured natural language rationales that reflect the influence of input features on model predictions. These rationales are then utilized to guide in-context learning, providing corrective signals that improve model performance without requiring human-annotated explanations. Their results show that post-hoc explanations can serve not only as mechanisms for understanding model behavior, but also as actionable components for diagnosing and improving reasoning and decision-making in LLMs, particularly in settings where explicit ground-truth rationales are unavailable. However, despite these advances, the framework does not rely on standardized or reference-based evaluation metrics to assess the fidelity of generated explanations, and it remains unclear whether these rationales align with human cognitive processes or simply optimize model performance.

Guo et al. [103] (2024) highlight key limitations of LLM-as-a-judge approaches, showing that such evaluators can exhibit biases and inconsistent judgments when assessing complex or open-ended outputs. Their findings motivate the integration of explainability and structured evaluation criteria to support more reliable and interpretable assessment of LLM behavior. However, while the study identifies critical limitations of LLM-based evaluation, it does not propose a quantitative and reproducible framework grounded in reference-based metrics, nor

does it assess whether the evaluation process aligns with human judgment and cognitive interpretation.

Wang et al. [104] (2024) proposed LLM-GAN and show its effectiveness in both prediction performance and explanation quality. To quantitatively evaluate explanation quality, they selected a set of quality metrics and used GPT-4o to score each explanation on a scale from 1 to 7. LLM-GAN outperformed GPT-4o in these metrics. The quality metrics include relevance to detection, fairness between real and fake news, accuracy, clarity, contextual understanding, and consistency. Their findings suggest that fine-tuned generative models can surpass general-purpose models in providing nuanced, structured evidence. Without a baseline of human expert evaluation (e.g., fact-checkers, journalists) to validate GPT-4o's scores, the evaluation remains somewhat insular. Research has shown that LLMs often exhibit egocentric bias (favoring text that mimics their own style).

Kim et al. [105] (2024) addressed the inherent "hallucination" risks in LLM explanations. They argued that a single LLM output is often prone to sycophancy (agreeing with the user) or unfounded claims. They proposed a multi-agent debate framework. In this setup, several LLM agents are assigned conflicting roles (proponent vs. opponent) to argue for or against a claim's veracity using external evidence. This approach involved using several LLMs that debate with each other, each agent providing arguments and counter-arguments to support or refute a claim. This highlights a critical need for consensus-based explainability rather than relying on a single generative path. However, the evaluation framework relies on LLM-based scoring, which introduces potential bias and lack of reproducibility, as highlighted in prior work on LLM-as-a-judge approaches. Furthermore, the absence of reference-based metrics or human-validated ground truth raises concerns about whether the generated explanations truly reflect faithful reasoning processes or merely optimize for the evaluation criteria.

Liu et al. [106] (2024) examined the effectiveness of using explanations to combat disinformation. The study looked at how the explanations generated by major LLM influence the perception of the credibility of information. Their study revealed that overly complex explanations can lead to cognitive overload, while overly simplistic ones fail to build trust. They conclude that XAI must be carefully designed to align with human mental models. The study does not propose a quantitative framework to measure the alignment between model-generated explanations and disinformation.

Boissonneault et al. [107] (2024) evaluated the performance of ChatGPT and Google Gemini for fake news detection using the LIAR dataset. LIAR contains 12,836 data manually classified by PolitiFact. Their results show that Google Gemini slightly outperformed ChatGPT, achieving an accuracy of 89.8%, a recall of 91.6%, a precision of 91.3%, and an F1-score of 91.4%. Although these findings show the strong potential of LLMs for fake news detection, the approach remains constrained by the need for task-specific tuning. However, the study does not propose a quantitative framework to measure the alignment between model-generated explanations and human cognitive responses, limiting the ability to systematically evaluate explanation effectiveness.

Sallami et al. [2] (2024) investigated the capacity of LLMs to detect fake news across three dimensions: (1) human-written fake news, (2) LLM-generated fake news, and (3) their comparative performance against a fine-tuned BERT model. This work represents the first systematic analysis of seven LLMs addressing both the production and detection aspects of disinformation. In their experiments, Gemma-1.1 achieved perfect accuracy, outperforming all other models. LLaMA-3 and Zephyr-ORPO also delivered strong results, each correctly identifying approximately 80% of the cases with very low error rates. In contrast, Mistral showed substantially lower effectiveness, detecting only around 20% of fake news instances. Additionally, the authors observed that LLM-generated fake news is significantly more chal-

lenging to detect than human-written disinformation, and that model scale generally correlates with improved detection performance. However, the reported performance varies significantly across models and may depend on the experimental setup and dataset characteristics, raising concerns about generalizability. Moreover, the study focuses primarily on detection accuracy and does not evaluate the quality, faithfulness, or interpretability of the explanations produced by these models, nor their alignment with human cognitive processes.

Sun et al. [1] (2024) conducted a comprehensive evaluation of fake news detection methods by combining automated model assessment with novel human-centered evaluation metrics. The authors compared multiple detection strategies, including fine-tuned pretrained language models, state-of-the-art fake news detection architectures, LoRA-fine-tuned LLMs, and ChatGPT-3.5 used in a zero-shot setting. The results indicate that fine-tuned Pretrained Language Models and specialized fake news detection models consistently outperform ChatGPT-3.5 when no fine-tuning is applied. However, LoRA-adapted LLMs show competitive performance while requiring significantly fewer trainable parameters. The human evaluation further revealed that although LLM-based predictions are often fluent and persuasive, they may lack consistency and factual grounding compared to supervised encoder-based models.

Taken together, these studies show rapid progress in leveraging LLMs for both fake news detection and explanation generation. The literature has explored diverse strategies, including chain-of-thought prompting, post-hoc rationale generation, multi-agent debate frameworks, and LLM-based judges. Despite these advances, explanation quality is most often assessed qualitatively or through subjective scoring by another LLM, which introduces potential bias and limits reproducibility.

Notably, none of the reviewed works employ reference-based semantic evaluation metrics such as BERTScore or COMET to quantitatively assess the quality and fidelity of LLM-generated explanations in fake news detection tasks. This limitation motivates the approach presented in Chapter 6 (Article 3), where explanation quality is evaluated using transformer-based similarity metrics grounded in human-written references, providing a reproducible and model-agnostic assessment framework.

In summary, while BERT and its variants dominated fake news detection research until approximately 2022, recent work has increasingly explored the use of large language models such as GPT, LLaMA, and Gemma. Nevertheless, encoder-based architectures remain highly relevant for fake news detection and related NLP classification tasks. **Pretrained encoder models continue to be widely adopted due to their computational efficiency, stability, and suitability for supervised learning scenarios where labeled datasets are available.** Unlike LLMs, BERT-style encoders generate compact contextual representations that can be fine-tuned with relatively modest computational resources, making them particularly well suited for large-scale, real-world fake news detection.

3.3. HUMAN COGNITIVE AND AFFECTIVE PERSPECTIVES

Recent advances in multimodal analysis have led to an increasing number of studies exploring the intersection of eye-tracking and fake news detection. These works aim to better understand how users visually process information and how cognitive effort varies when individuals are exposed to misleading content.

Early contributions in this area focused primarily on global eye-tracking metrics. For instance, Sümer et al. [22] (2021) collected data from 25 participants using 60 news items and proposed the FakeNewsPerception dataset. The study measured total viewing time, number of

fixations, and number of saccades, with believability rated on a 5-point Likert scale. Their results show that processing fake news induces a higher cognitive load, characterized by significantly higher fixation counts and longer reading times compared to real news. However, this approach masks the specific textual features that demand this extra processing. Our study aims to disaggregate this cognitive load, hypothesizing that the increased fixation counts observed by Sümer et al. are not evenly distributed, but are densely clustered around PPE.

Davila et al. [25] combined eye-tracking with emotion recognition techniques to investigate user responses to political fake news. Their study identified significant correlations between emotional facial expressions and exposure to false claims, suggesting that emotional reactions play a key role in the processing and spread of misinformation. However, their approach relied primarily on facial emotion recognition, which captures observable behavioral responses rather than internal cognitive or affective states. As such, it may not fully reflect the underlying neural mechanisms involved in processing deceptive information. In addition, facial expressions can be influenced by social context, individual differences, or conscious control, which may limit the reliability of these signals as indicators of genuine emotional engagement. In contrast, physiological measures such as EEG provide a more direct and objective assessment of neural activity, enabling the detection of subtle cognitive and affective processes that are not externally visible. Therefore, while Davila et al.'s work highlights the importance of emotional responses in fake news perception, it also underscores the need for integrating more direct neurophysiological measurements to better understand the underlying mechanisms of user susceptibility to misinformation.

Building on this dataset, Bozkir et al. [108] (2022) investigated regressive saccadic eye movements as an indicator of cognitive difficulty. Their results revealed that participants exhibited significantly more regressive saccadic eye movements when reading fake news (M=4.28, SD=2.96) than on the real news content (M=3.85, SD=2.47) with (p=.007, U=

281,809 and CLES=.54). They used Mann-Whitney U test with level of significance α=.05 and brute-force version of Common Language Effect Size (CLES) statistic. While Bozkir et al. established that readers regress more frequently when processing fake news, they did not investigate what specific elements triggered these backward movements. We hypothesized that these regressive saccades were not random, but are driven by the need to verify specific semantic anchors. Our study extends this line of research by identifying PPE as the primary targets of this increased visual attention and verification effort.

Shi et al. [24] (2023) used eye-tracking with 35 participants to investigate how COVID-19 fake news affects cognitive activity. They conducted a one-way ANOVA ($F(4, 3004)=321.9$, $p<.05$) followed by post hoc Tukey's HSD tests. They also performed a two-way ANOVA examining the effects of headline stance and claim correctness on relative pupil dilation (RPD), pupil diameter, fixation frequency, fixation duration, and cognitive load. The analyses revealed significant main effects for both claim correctness ($F(1,34)=15.54$, $p<.05$) and headline stance ($F(1,34)=31.98$, $p<.05$). Their statistical analyses revealed that both claim correctness and headline stance significantly influence cognitive load, as reflected in pupil dilation and fixation behavior. These findings suggest that the mental effort required to evaluate information depends not only on its veracity but also on how it is framed. Our study addresses this gap by investigating how PPE specifically drive visual attention and cognitive processing in fake news.

Steinfeld [109] (2023) investigated eye movements in a study where 83 participants were asked to assess the credibility of online messages posted on social media and web pages. The study employed multivariate regression analysis and measured ocular attention to metadata areas, gaze patterns, and dwelling time on webpage areas that provide credibility cues. The study took into account the website URL, message source, username, and profile information on a social media page, as well as comments made by other users in response to the message.

All of the insights made successful identification of fake news positively linked to ocular attention to information metadata. This study focused on eye-tracking movement to analyze the different parts (areas of interest) of the Web pages that participants consulted before detecting fake news. While Steinfeld focused on external context cues (metadata and layout), our study shifts the focus to internal content cues embedded within the text itself. Specifically, we investigated how visual attention is allocated to PPE, hypothesizing that readers verify the content by focusing on the specific actors and organizations mentioned, rather than relying solely on the source's reputation.

De Filippi et al. [110] (2024) presented a proof-of-concept study exploring multimodal for misinformation detection through the analysis of eye-tracking and physiological responses (galvanic skin response, heart rate variability, pupil dilation) to capture unconscious behavioral and biological responses during disinformation exposure. The results show superior performance of multimodal classification compared to unimodal approaches for distinguishing true from false news, highlighting the advantage of integrating different behavioral metrics in detection systems based on readers' cognitive and emotional responses. This work complements existing EEG-based approaches by offering additional physiological modalities for comprehensive detection systems.

Lutz et al. [111] (2024) conducted a within-subject study with 42 participants evaluating 40 news articles (20 real and 20 false). They measured cognitive processing through eye fixations and affective processing through heart rate variability. They performed a regression analysis to explain cognitive and affective information processing based on different linguistic cues (adverbs, personal pronouns, positive emotion words, negative emotion words). The linguistic cues were calculated with the "Linguistic Inquiry and Word Count" tool and represented a set of words grouped along different dimensions. Given the within-subject design, they performed a mixed-effects regression analysis to account for individual variances.

This allowed them to model the specific impact of linguistic cues on cognitive and affective processing while controlling for participant baselines. They found that structural features like article *length* and analytic style (formal language) influence cognitive and affective processing. However, their study focused primarily on stylistic and psychometric linguistic categories rather than specific semantic entities. In contrast, our paper focuses on PPE as distinct semantic anchors.

Taken together, these studies show that fake news is associated with distinct patterns of visual attention and increased cognitive load, including longer fixations, more regressions, and greater pupil dilation [108]. However, most prior studies analyze these effects at a global level, focusing on aggregated measures such as total viewing time or overall fixation counts. As a result, they provide limited insight into the specific textual elements that trigger these cognitive responses.

In particular, while previous research has examined stylistic cues, source metadata, and layout-level features, little attention has been paid to the role of semantic anchors embedded within the text itself. The literature does not yet clarify which categories of words or entities systematically attract visual attention during fake news evaluation, nor how such attention relates to heuristic credibility judgments. To address this gap, the contribution appeared in Chapter 6 addresses this limitation by focusing on proper nouns of political figures as potential cognitive anchors, using eye-tracking data to analyze how these elements shape visual attention and influence veracity judgments at the individual reader level.

3.3. CHAPTER SUMMARY

The analysis of the literature highlights specific limitations inherent to each approach. Taken together, these shortcomings underscore a critical need for an integrated framework

that bridges fake news detection, interpretability for humans, and human cognition. To address this challenge, the next chapter introduces the first contribution, which leverages Transformers (specifically BERT) and Explainable AI (XAI) to improve fake news detection and explainability.

CHAPTER IV

BERT AND XAI FOR MITIGATING FAKE NEWS

4.1. INTRODUCTION

This chapter presents the methodology developed in our first contribution [112] to evaluate how contextual metadata can enhance BERT and LSTM models for fake news detection. It also introduces SHAP which quantifies the contribution of each metadata attribute to the final prediction. Together, these elements establish the methodological foundation for assessing whether contextual information improves both performance and explainability in fake news detection.

4.2. RESEARCH QUESTIONS

Several studies have shown that BERT-based architectures achieve strong performance in fake news detection tasks [87, 50]. However, in most existing approaches, contextual information is not directly integrated into the transformer architecture. Instead, it is often processed through separate modules or concatenated as external features, which limits the model's ability to jointly learn interactions between textual and contextual signals within a unified representation space.

From an explainability perspective, post-hoc explanations often lack semantic consistency with the underlying decision process. In particular, keyword-based explanations frequently fail to reflect the deeper contextual mechanisms driving model predictions, suggesting that current XAI approaches may not fully capture the reasoning patterns of transformer-based classifiers in fake news detection settings.

To address these limitations, this contribution proposes a unified BERT-based framework that incorporates contextual metadata directly into the input representation. In addition, SHapley Additive exPlanations (SHAP) is employed as a post-hoc explainable method to quantify the marginal contribution of each contextual attribute to the model's predictions, based on cooperative game-theoretic principles.

This contribution investigates the two research questions derived from sub-hypothesis 1:

RQ1.1: Does the integration of contextual metadata improve the performance of deep learning models for fake news detection?

RQ1.2: Which contextual metadata features contribute most significantly to fake news detection?

Sub-hypothesis 1: Augmenting a BERT-based architecture with contextual metadata (e.g., number of followers, and hashtags), combined with SHAP-based feature attribution, improves both predictive fake news performance and explainability.

4.3. BACKGROUND

This section introduces the key background concepts underlying fake news detection. It begins with content-based approaches focusing on linguistic features, then explores context-based methods that incorporate metadata and social signals. It further examines the mechanisms of information propagation and concludes with an overview of existing countermeasures aimed at limiting the spread of fake news.

4.3.1. CONTENT-BASED APPROACH

Content-based fake news detection focuses on analyzing the textual content of news articles, including the body text, headlines, and sometimes associated user comments. Early approaches relied on statistical representations such as Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF), which model text as unordered collections of words. Although these methods were effective for simple categorization tasks, they are now considered insufficient for fake news detection because they ignore word order, long-range dependencies, and contextual meaning, elements that are essential for identifying subtle manipulative cues.

Modern content-based approaches instead rely on deep linguistic features, which capture progressively richer levels of language structure. These features operate across three conceptual layers that reflect how natural language conveys meaning. **At the lexical level**, the analysis focuses on individual words and their surface properties. Early studies, such as Shu et al. [20], extracted features including word frequency distributions, character and word n-grams, sentiment lexicons, and psycholinguistic indicators derived from tools like LIWC. These features help identify stylistic markers commonly associated with fake news, such as emotionally charged vocabulary, exaggerated adjectives, or informal language intended to provoke strong reactions.

The syntactic level examines how words are arranged into grammatical structures. Research has shown that fabricated content often exhibits distinctive syntactic signatures compared to professionally written journalism. These include simpler or irregular sentence constructions, an overuse of superlatives, unconventional punctuation, and atypical part-of-speech patterns. Shu et al. [20] show that such syntactic irregularities can serve as reliable indicators

of deceptive writing, as they reflect the rhetorical strategies used to attract attention or create urgency.

The semantic level represents the deepest layer of linguistic analysis and focuses on the meaning conveyed by the text. Semantic features capture relationships between concepts, topic coherence, and the alignment between different parts of a document. This level is particularly important for detecting sophisticated forms of fake news, where the deception lies not in the vocabulary itself but in subtle inconsistencies, misleading framing, or contradictions between the headline and the body text. The transition, from statistical models to transformer-based architectures, has revolutionized semantic analysis. Through self-attention mechanisms, these models learn to weigh the importance of specific words and phrases within the global context of a sentence or document, enabling them to detect nuanced linguistic signals that earlier models could not capture.

Several studies illustrate how these deep linguistic features are operationalized in practice. In 2017, Shu et al. [20] combined lexical, syntactic, and semantic indicators to build a comprehensive linguistic profile of fake news. Their work incorporated sentiment polarity, readability metrics, part-of-speech distributions, and topic modeling to capture both surface-level and conceptual characteristics of deceptive content. In 2019, Shu et al. [7] extended this approach by analyzing semantic similarity between headlines and body text, as well as rhetorical structures and stance alignment. In 2020, they [85] further introduced temporal linguistic drift, showing that the wording of fake news evolves over time and that early detection requires capturing these dynamic patterns.

Despite their sophistication, content-based approaches face several limitations. The linguistic style of fake news evolves rapidly, making models trained on older data less effective when confronted with new narratives or emerging topics. Domain shift is another

challenge: a classifier trained on political fake news may perform poorly on health-related fake news because the linguistic cues differ significantly. Moreover, many linguistic features are language-specific, limiting the generalizability of models across languages. Short texts, such as tweets, often lack sufficient linguistic content for reliable detection, and deep learning models require large annotated datasets that are costly and time-consuming to produce. As Raza et al. [113] noted, by the time enough labeled data is collected, fake news has often already spread widely. These limitations highlight a fundamental issue: text alone is often insufficient to detect fake news, especially when the content is crafted to appear credible. This motivates the integration of contextual information such as user metadata, network structure, and temporal dynamics which forms the basis of the context-based and multimodal approaches explored in the subsequent sections of this thesis.

4.3.2. CONTEXT-BASED APPROACH

While content-based approaches focus exclusively on the linguistic properties of a message, context-based fake news detection incorporates metadata that describe how, when, and by whom information is produced and disseminated. This perspective is grounded in the observation that fake news is not merely a textual artifact but a social phenomenon shaped by user behavior, network structure, and temporal dynamics. As a result, context-based approaches aim to capture the broader ecosystem within which fake news circulates, providing insights that are often inaccessible through textual analysis alone.

Metadata refers to auxiliary information associated with a post that does not include its textual or visual content. In social media environments, metadata plays a crucial role in understanding dissemination patterns, identifying anomalous behaviors, and modeling the underlying social dynamics of information diffusion. Shu et al. [7] described the importance of contextual signals through their tri-relationship model, which jointly analyzes publishers,

news articles, and user interactions. Their work showed that contextual cues such as user credibility, propagation structure, and early engagement can significantly improve fake news detection, particularly in early-stage scenarios where textual cues may be insufficient.

A first category of contextual information is **content-based metadata**, which captures structural and descriptive attributes of a post. These include the timestamp of publication, language, media type (text, image, video, hyperlink), message length, and the presence of hashtags, mentions, or URLs. Such features provide valuable information about how content is formatted and disseminated. For example, fake news posts often rely on short, emotionally charged messages accompanied by external links or sensational hashtags. Temporal metadata embedded in these attributes also helps identify unusual posting patterns, such as bursts of activity at atypical hours or synchronized publication across multiple accounts.

A second category is **user-based metadata**, which characterizes the accounts responsible for generating or sharing content. Features such as the number of followers and followings, account age, posting frequency, and verification status provide insight into user credibility and influence. These indicators are particularly useful for detecting automated accounts (bots) and coordinated campaigns. Research has shown that bots often exhibit extreme posting frequencies, highly imbalanced follower-following ratios, or recently created accounts designed to amplify specific narratives. Shu et al. [85] emphasized that user-level features are essential for early detection, as fake news often originates from low-credibility or newly created accounts before spreading widely.

A third dimension is **interaction metadata**, which reflects the engagement generated by a piece of content. This includes the number of likes, shares, and comments. Interaction patterns can reveal whether a post is spreading organically or being artificially amplified. The

structure of discussion threads also provides insight into how users react to and propagate the content.

A fourth category is **network metadata**, which describes the relational structure between users. This includes follower-following relationships, retweet and mention networks, and the formation of communities or clusters. Network analysis enables the identification of diffusion cascades, echo chambers, and coordinated behavior patterns. Graph-based approaches model these relationships to detect anomalies such as tightly connected groups repeatedly sharing the same content (echo chambers) or accounts that act as central hubs in the propagation of fake news. These structural insights are essential for understanding how fake news spreads within and across communities, often bypassing traditional fact-checking mechanisms.

Finally, **temporal metadata** captures the dynamic evolution of information diffusion over time. Beyond simple timestamps, temporal features include inter-posting intervals, frequency bursts, and the overall shape of propagation curves. Fake news often exhibits "burstiness," characterized by rapid and concentrated diffusion followed by irregular secondary peaks. Shu et al. [85] showed that these temporal signatures can be leveraged for early detection, as fake news tends to spread more aggressively and unpredictably than factual content.

The integration of metadata into fake news detection offers several advantages. Vosoughi et al. [6] showed that fake news spreads "farther, faster, and deeper" than factual information, highlighting the importance of modeling engagement dynamics. It enables early detection without requiring deep semantic analysis, provides robustness against linguistic manipulation, and can be incorporated into hybrid models such as Graph Neural Networks (GNNs). Metadata also complements textual features by capturing behavioral and structural cues that are difficult to manipulate consistently. However, several limitations must be acknowledged.

Metadata availability is platform-dependent and often restricted due to privacy policies and API limitations. There is also a risk of conflating correlation with causation; for example, high virality does not necessarily imply falsity. Moreover, adversarial actors may deliberately manipulate metadata signals, for instance, by using sophisticated bots that mimic human behavior making detection more challenging.

Overall, integrating various algorithmic approaches, which address content and context, can enhance the detection of fake news [87, 13]. This perspective forms the foundation for hybrid models that combine textual analysis with metadata, such as the BERT-based architecture which integrates text and metadata features to improve detection performance and explainability.

4.3.3. VECTORS OF PROPAGATION

The mechanisms supporting rapid diffusion are diverse and evolving. Based on an analysis conducted between 2013 and 2018, Martin and Shapiro [114] identified the most common dissemination techniques used by hostile actors as, in decreasing order of importance: **trolls and bots, fake accounts, hijacking social accounts, hashtags, and deepfakes**. Recent reports from the Canadian Centre for Cybersecurity [115] and the Hogue Commission [116] indicate that the hierarchy has shifted significantly with the rise of generative AI. Automated bots are increasingly replaced by AI agents capable of sustaining complex conversations, adapting arguments in real time, and feigning empathy. Simple fake accounts with a stolen photo have been replaced by complete synthetic identities generated by Generative Adversarial Network (GAN). Deepfakes, ranked 4 in 2018, became in 2026 a leading weapon.

Algorithmic recommendation systems further exacerbate the spread of fake news. Social media algorithms, driven by artificial intelligence, prioritize content based on user

engagement signals such as likes, shares, and reactions. Fake news often benefits from these mechanisms due to its emotional appeal and plausibility, prompting users to engage with and redistribute content they perceive as credible. This engagement feeds algorithmic amplification, reinforcing echo chambers in which users are repeatedly exposed to similar narratives [117].

Fake accounts and coordinated inauthentic behaviour further distort perceptions of consensus. During the 2016 US election, the Twitter account @TEN_GOP, falsely presented as the official account of the Tennessee Republican Party, amassed hundreds of thousands of followers and generated substantial engagement [117]. In 2018, the US Department of Justice accused thirteen Russians of interfering in the US 2016 presidential election, plus three Russian entities, including the IRA. The purpose of the charges is based on stolen identities and fake social media accounts. This fake operation was backed by a Kremlin associate, Prigozhin, known as “the chef” who is also accused [32].

Trolls can be defined as a human that aims to generate controversy, unlike **bots** which are robots that emulate the activity of human users. Cyber threat actors can create a false impression that millions of people share the message using automated tools and techniques, like bots and trolls. Diffusion methods have mutated. The study by Martin and Shapiro (2019) [114] indicates that "bot" techniques are utilized 90% of the time. Bots, in conjunction with hashtag hijacking, are identified as the primary methods for disseminating fake news. Nowadays, repetitive bots have been replaced by agents powered by Large Language Models (LLMs), capable of sustaining nuanced and undetectable conversations. Moreover, in 2017, it is estimated that bots account for 9 to 15% of all Twitter accounts [118].

Hashtags have also played a critical role in accelerating fake news diffusion. A notable instance is the use of the hashtag BlackLivesMatter, which has evolved into a symbol of racial injustice [119]. As noted by Sinan Aral [120], a simple interaction by a public figure or a bot

with a hashtag can trigger a chain reaction, transforming a local rumor into a global trend. For example, Elon Musk has more than 40 million Twitter followers and by adding #bitcoin to his post and tweeting about dogecoin, he has increased cryptocurrency market events [120].

4.3.4. MEASURES TO COUNTERFEIT

Several measures are being implemented to combat the spread of fake news including media literacy initiatives and platform-based interventions. Line Pagé, the founder and president of the Board of Directors of the Quebec Center for Media and Information Education (CQEMI), emphasizes the necessity of educating young people, as over 90% encounter fake news regularly in their daily lives [121]. The CQEMI offers training to its affiliated schools.

Social networks have also implemented measures to combat fake news. They employ strategies such as tagging and redirecting to curb the dissemination of false news. Redirecting involves linking content to its official site, while tagging entails adding a label, such as "!" to prompt users to get verified information on topics like COVID-19. These measures appear ineffective in halting the spread of fake news. On one hand, there is a necessity to discover a rapid and automated method for its detection to prevent detrimental effects. On the other hand, employing the most recent advancements in artificial intelligence could enhance our comprehension of the different facets of fake news and elucidate the processes involved in its creation and detection.

Recent developments further complicate the landscape. In January 2025, Meta announced the end of its third-party fact-checking program in the United States, opting for a community-based moderation model similar to that of platform X [11]. This decision has drawn criticism, with some arguing that it could facilitate the spread of fake news on X.

Faced with this threat, legislators attempt to build defenses, but regulation struggles to keep pace with technological evolution.

The Europe framework is the most advanced. While the 2018 French national law against information manipulation had a limited impact, proving falsehood within 48 hours being legally arduous, it paved the way for systemic regulation. Since then, the European Union has shifted the paradigm with the **Digital Services Act** (DSA), fully effective as of February 2024. Unlike the French law which targeted specific content, the DSA regulates the platforms themselves, forcing "Gatekeepers" (defined under the parallel **Digital Markets Act**) to mitigate systemic risks like disinformation or face massive fines [122].

In the United States, the response is fragmented across the states. For example, some states (e.g., California, Texas) are legislating urgently, particularly to address the threat of electoral deepfakes.

In Canada, Bill C-70, adopted in June 2024, modernizes the fight against foreign interference and creates a transparency registry. However, the regulation of hateful or false content on platforms remains in limbo, creating a vulnerability gap. Paradoxically, some platforms are retreating: in January 2025, Meta scaled back its third-party fact-checking program, opting for a controversial community moderation model [11].

4.4. METHODOLOGY

The proposed methodology follows a multi-stage pipeline consisting of data collection, preprocessing, model development, training, evaluation, and explainability analysis. The contribution investigates the impact of integrating contextual metadata into deep learning models for fake news detection and evaluates BERT and BiLSTM models.

4.4.1. DATA COLLECTION

We used a publicly available COVID-19 fake news dataset extracted from the Twitter platform. The dataset contains both textual content and contextual metadata (12 features) associated with each post. From the full corpus, we selected the first 11,000 entries. Table 4.1 summarizes the metadata fields included in the analysis.

Table 4.1 : Metadata fields extracted from the COVID-19 dataset

Feature	Description
User location	City or region provided by the user
User description	User profile biography
Account creation date	Timestamp of account creation
Number of followers	Size of the user's audience
Number of friends	Number of accounts followed
Number of favorites	Number of liked tweets
Verified status	Whether the account is verified
Publication date	Timestamp of the tweet
Text with URL	Tweet text including hyperlinks
Hashtags	Topical markers included in the post
Source	Data extracted from Twitter API
Retweet flag	Indicates whether the post is a retweet

4.4.2. DATA PREPROCESSING

Data preprocessing was performed on both textual and metadata features. Hyperlinks were removed using regular expressions. No stemming or stopword removal was applied, as BERT relies on subword tokenization that preserves contextual semantics. After the preprocessing, 5,000 data remained.

Metadata features were processed using different strategies depending on their type:

- Numerical features (followers, friends, favorites) were log-transformed using $\log(1+x)$ and standardized.
- Categorical features (verified status, user location, source) were encoded using learned embedding representations.
- Temporal features (account creation date, publication date) were transformed into derived variables such as account age and posting recency.

Both the BERT-based and BiLSTM models are trained using the same feature set, consisting of textual content and contextual metadata. This ensures a fair comparison between architectures.

4.4.3. BERT MODEL ARCHITECTURE

The proposed BERT-based architecture is a multimodal model that integrates textual and metadata representations for fake news classification. As shown in Figure 4.1, the textual input is tokenized and passed to a BERT-base encoder. The contextual representation is extracted from the [CLS] token of the final hidden layer. In parallel, numerical, categorical, and temporal metadata are embedded and projected into dense representations. These include log-transformed numerical signals (followers, friends, favorites), categorical embeddings (user location, verified status, source), and engineered temporal features.

The final representation is obtained by concatenation:

$$h_{\text{joint}} = [h_{\text{CLS}}; e_{\text{num}}; e_{\text{cat}}; e_{\text{temp}}] \tag{4.1}$$

This vector is passed through a dense layer with ReLU activation, followed by dropout regularization and a sigmoid activation function for binary classification. Table 4.2 summarizes all the hyperparameters used for BERT.

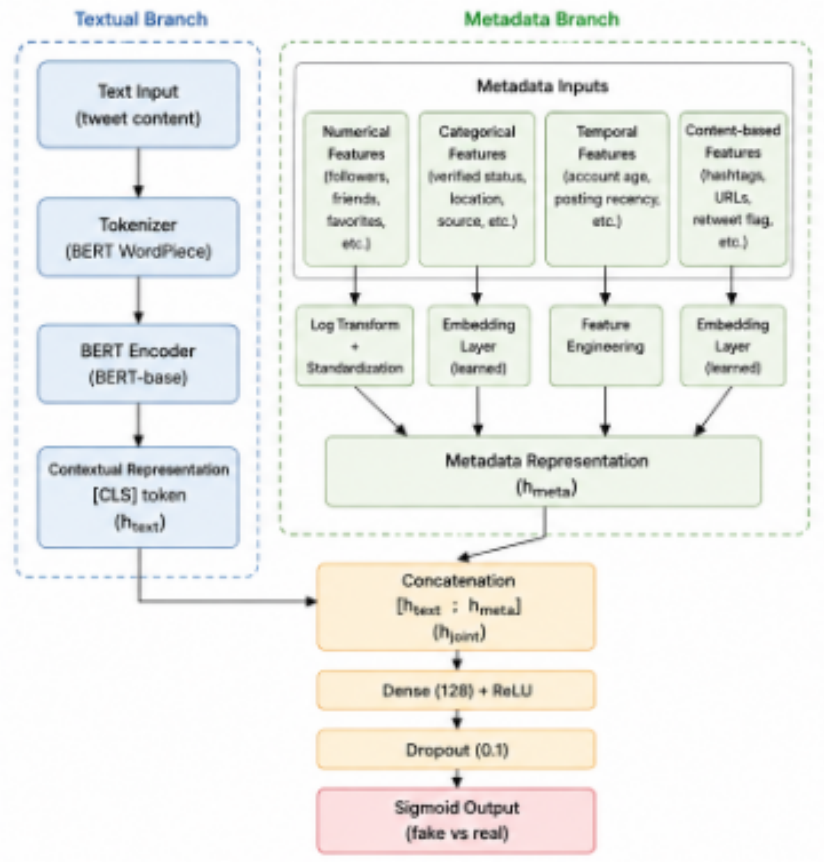


Figure 4.1 : BERT-based Architecture

Table 4.2 : BERT Hyperparameters

Hyperparameters	Value
Model	BERT-base (12 layers, 12 heads, hidden size 768)
Max sequence length	512
Batch size	16
Learning rate	2×10^{-5}
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.01)
Warmup ratio	0.1
Dropout	0.1
Gradient clipping	1.0
Epochs	4
Early stopping	Patience = 2
Loss function	Cross-entropy

4.4.4. LSTM MODEL ARCHITECTURE

The bidirectional LSTM (BiLSTM) model was implemented to compare sequential learning performance against transformer-based architecture. As illustrated in Figure 4.2, the BiLSTM baseline model processes both textual content and contextual metadata using pretrained GloVe embeddings for text and embedded representations for metadata features. The BiLSTM layer captures sequential dependencies across the combined feature representation, followed by dense layers for classification.

The model uses 300-dimensional GloVe embeddings as input representations. A BiLSTM layer with 256 hidden units is applied to capture contextual dependencies. The output is passed through a fully connected layer with 128 units and ReLU activation, followed by dropout and a sigmoid classification layer.

For the context-enhanced configuration, the BiLSTM textual representation was concatenated with processed metadata features, including numerical, categorical, and temporal

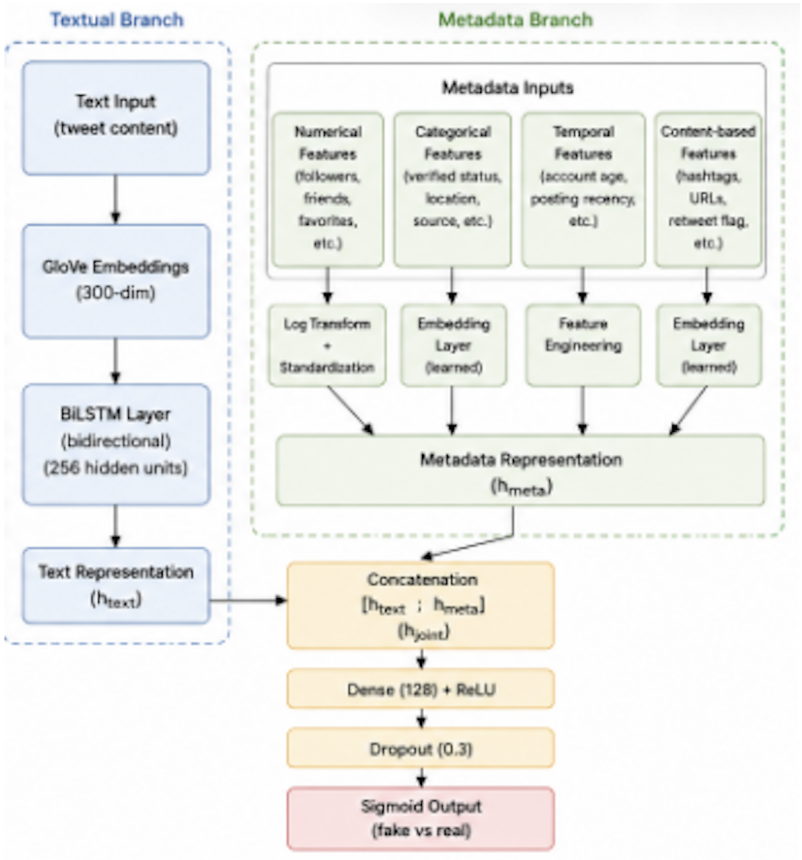


Figure 4.2 : LSTM Architecture

attributes, similarly to the multimodal BERT architecture. The resulting joint representation was then passed through the dense classification layers to evaluate the impact of contextual information on recurrent architectures. GloVe embeddings were selected due to their ability to capture global word co-occurrence statistics, providing rich semantic representations while remaining computationally efficient compared to contextual embeddings. The BiLSTM allows the model to capture both past and future contextual dependencies, which is particularly important for understanding the semantic structure of sentences. Table 4.3 summarizes all the hyperparameters used for BiLSTM.

Table 4.3 : BiLSTM Hyperparameters

Hyperparameter	**Value**
Embedding dimension	300 (GloVe pretrained)
LSTM architecture	Bidirectional LSTM
Hidden size	256
Dropout	0.3
Fully connected layer	128 units + ReLU
Batch size	32
Learning rate	1×10^{-3}
Optimizer	Adam
Gradient clipping	5.0
Epochs	10
Early stopping	Patience = 3
Loss function	Binary cross-entropy

4.4.5. TRAINING PROCEDURE

The dataset was split into 80% training and 20% testing. For the BERT-based model, the AdamW optimizer was used with a learning rate of 2×10^{-5}, batch size of 16, and 4 training epochs. Early stopping was applied based on validation loss. For the BiLSTM model, the Adam optimizer was used with a learning rate of 1×10^{-3}, batch size of 32, and gradient clipping at 5.0. All experiments were conducted using Google Colab with a T4 GPU.

4.4.6. EVALUATION METRICS

Fake news detection primarily relies on supervised learning methods, using labeled datasets that contain news articles already classified as real or fake. This allows us to compare the model' classifications against the ground truth in the dataset, enabling us to analyze their performance. Input datasets must be annotated as either true or false, and ideally balanced so that the model is exposed to an equal number of examples from each class.

To evaluate the effectiveness of fake news detection, we used the following metrics. Note that the class we care most about detecting is fake news. Therefore, fake news is the positive element and true news is the negative one. Table 4.4 presents the formula for both, false or true news. The definitions for each metric are as follows:

- Accuracy: proportion of correct predictions over all predictions.
- Precision: proportion of correctly predicted positive samples among all predicted positives.
- Recall: proportion of correctly predicted positive samples among all actual positives.
- F1-score: harmonic mean of precision and recall.

Table 4.4 : Performance Formula

Metric	Fake news	True news
Precision	TP/(TP+FP)	TN/(TN+FN)
Recall	TP/(TP+FN)	TN/(TN+FP)
F1-score $2 \times (P \times R)/(P+R)$		
Accuracy $(TP+TN)/(TP+FP+TN+FN)$		

Note: TP: True Positive, FP: False Positive, TN: True Negative, FN: False negative.

4.4.7. SHAPLEY VALUES

Explainability techniques such as SHAP (SHapley Additive exPlanations) were used to analyze feature importance and provide post-hoc explainability of the proposed models. To explain the model's predictions, we applied Shap.Explainer framework, which provides a unified interface for estimating Shapley values across different model types. The same tokenizer and preprocessing pipeline used during training were applied to ensure consistency. A background dataset of 200 samples was used to estimate Shapley values. SHAP was used to estimate the contribution of textual tokens and metadata features to individual predictions.

4.4.8. LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME)

To extract local feature importance for each news claim, we implemented the LIME (Local Interpretable Model-agnostic Explanations) framework using the LimeTextExplainer module. The implementation was carried out in a Google Colab environment.

The LimeTextExplainer generates local surrogate explanations by approximating the model's behavior around a given prediction. It assigns importance weights to individual words, providing an approximation of the textual features that influenced the classification outcome.

4.5. RESULTS

RQ1.1: Does the integration of contextual metadata improve the performance of deep learning models for fake news detection?

BERT has shown excellent performance in detecting fake news using textual content alone. When reviewing prior studies that applied deep learning to fake news detection [94, 63], we observed that BERT was typically used for content analysis, whereas context features, when included, were processed using separate algorithms rather than being integrated directly into BERT.

Based on Table 4.5, BERT achieved an accuracy of 97.16% for content-only analysis, compared to 94.79% for LSTM. This indicates that BERT outperforms LSTM by 2.37% in detecting fake news using textual features alone. When context features were added, BERT reached an accuracy of 99.84%, while LSTM obtained 97.60%, meaning that BERT still performed better by 2.24%.

These results suggest that incorporating context features improves performance for both models, but BERT consistently achieves higher accuracy, confirming its superior ability to leverage both content and contextual information for fake news detection.

Table 4.5 : BERT and LSTM results

Classifier	Approach	Accuracy	Precision	F-score	Recall
BERT	Text-only	97.16	97.02	97.01	96.88
LSTM	Text-only	94.79	94.30	94.10	93.30
BERT	Text+metadata	99.84	99.00	99.00	99.00
LSTM	Text+metadata	97.60	97.00	97.00	97.00

We can conclude that adding context features improves fake news detection by 2.68%, increasing accuracy from 97.16% to 99.84%. This confirms previous findings showing that BERT outperforms other deep learning models in fake news detection tasks [113].

RQ1.2: Which contextual metadata features contribute most significantly to fake news detection?

SHAP provided more meaningful and consistent explanations compared to LIME. Figure 4.3 presents the SHAP results and highlights the most influential contextual features contributing to fake news detection.

Among all features, *user_followers* and *hashtags* emerge as the most significant. Specifically, *user_followers* contributes approximately 25% to the model's predictions, while *hashtags* account for around 20%. Taken together, these two features represent nearly half of the total explanatory power (approximately 45%), indicating that they play a dominant role in detecting fake news.

Furthermore, our analysis reveals that posts containing a higher number of hashtags are more likely to be associated with fake news. As a result, SHAP assigns a positive contribution to the *hashtag_count* feature when predicting fake content.

In addition to these dominant features, several other attributes contribute moderately to the classification task. These include *user_location*, *user_friends*, *user_verified*, and *data_text*, each accounting for approximately 10% of the explanatory contribution.

In contrast, features such as *user_favourites*, *user_created*, and *user_description* exhibit minimal influence on the model's predictions, suggesting that they play a limited role in distinguishing between fake and real news.

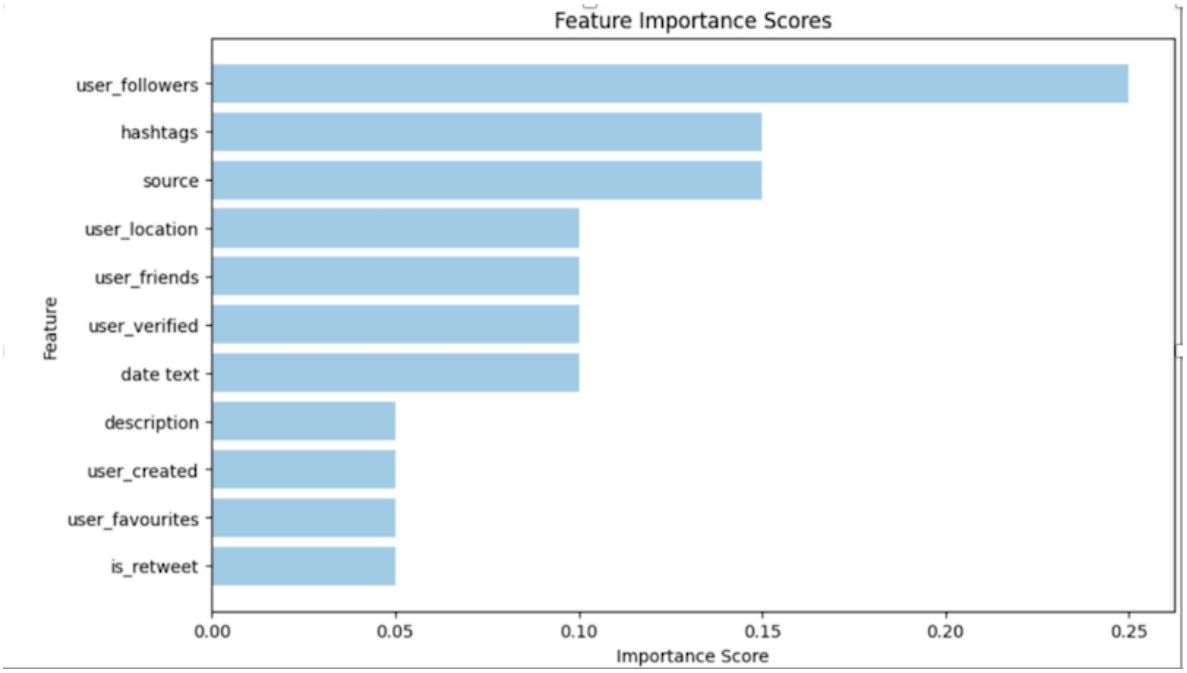


Figure 4.3 : SHAP result

4.6. DISCUSSION

The results obtained in this contribution presents that integrating contextual metadata into a BERT-based architecture significantly enhances fake news detection performance. This finding aligns with recent research showing that linguistic signals alone are often insufficient to capture the behavioural and structural patterns that characterize deceptive content. In this section, we discuss the implications of our results, the role of metadata, the contribution of explainability, and the limitations of the proposed approach.

4.6.1. WHY BERT PERFORMS WELL ON FAKE NEWS DETECTION

BERT's strong performance can be attributed to several architectural properties. First, its large-scale pretraining on billions of words enables the model to encode rich semantic and syntactic knowledge. Second, the Transformer architecture allows BERT to capture long-range dependencies through multi-head self-attention, which is essential for identifying subtle inconsistencies or rhetorical cues often present in fake news. Third, the masked language modeling objective forces the model to develop deep contextual understanding rather than relying on surface-level patterns. Finally, BERT's bidirectional encoding provides a holistic view of the text, allowing it to detect contradictions or misleading framing that may span multiple parts of a message.

These properties explain why BERT consistently outperforms classical machine learning models and earlier deep learning architectures in fake news detection tasks. Prior studies have shown that BERT-based models achieve state-of-the-art performance across multiple COVID-19 misinformation datasets [95]. More recent work confirms that fine-tuning is essential for optimal performance: a 2025 comparative study found that BERT, RoBERTa, DistilBERT, ALBERT, and BERTweet all require task-specific fine-tuning to reach their full potential [123]. Our results are consistent with these findings.

4.6.2. CONTRIBUTION OF METADATA TO MODEL PERFORMANCE

While BERT effectively captures linguistic cues, fake news is not solely a textual phenomenon. It is also shaped by user behaviour, network structure, and temporal dynamics. Metadata therefore provides complementary information that helps the model distinguish between credible and non-credible content.

SHAP analysis revealed that several metadata features played a significant role in the model's predictions:

User followers. The number of followers reflects the potential influence of an account. Although high-follower accounts can amplify fake news, prior research shows that most individuals spreading false information are humans rather than bots [6]. SHAP values indicate that follower count contributes to the model's assessment of credibility, particularly when combined with other behavioural signals.

Hashtags. Hashtags are powerful indicators of topical alignment and coordinated behaviour. Prior studies show that trolls and coordinated groups exploit popular hashtags to increase the visibility of fake news [88]. In our dataset, the number of hashtags, not their semantic content, was a strong predictor of fake news, consistent with findings that COVID-19 conspiracy hashtags are disproportionately associated with fake news [124].

User location. Location metadata can reveal inconsistencies between the content of a message and the user's reported geographical context. It can also help identify geographically targeted disinformation campaigns. SHAP analysis showed that location embeddings contributed meaningfully to the model's decisions.

User friends. The number of friends reflects the structure of a user's social circle. Echo chambers, tight-knit communities where users reinforce each other's beliefs, play a central role in the spread of fake news. The user_friends feature helped the model capture these propagation patterns.

Overall, metadata embeddings enriched the [CLS] representation by providing behavioural and contextual cues that are not present in the text alone. This confirms our hypothesis that hybrid architectures combining content and context outperform text-only models.

4.6.3. ROLE OF EXPLAINABILITY

The integration of SHAP (SHapley Additive exPlanations) provides a post-hoc explainability framework to quantify feature-level contributions to individual predictions. In contrast to attention-based heuristics, which have been shown to exhibit weak and often inconsistent correlations with feature importance [49], SHAP values are grounded in cooperative game theory and satisfy desirable properties such as local accuracy and consistency.

Within this framework, SHAP enables a decomposition of the model output into additive contributions attributed to each input feature, thereby offering a principled mechanism for explainability analysis. This approach was used to quantify the relative importance of metadata features in the classification decision, to assess the extent to which predictions are driven by potentially spurious correlations, to examine the interaction between textual and metadata signals in the learned representations, and to perform qualitative error analysis through token- and feature-level attribution patterns.

From a methodological standpoint, explainability is a critical requirement in fake news detection, particularly in high-stakes applications where transparency, auditability, and user trust are essential. Moreover, it aligns with emerging requirements in algorithmic accountability and regulatory compliance frameworks governing automated decision-making systems.

4.7. THREATS TO VALIDITY

Our methodology has some limits that has to be addressed.

Construct validity: Construct validity concerns whether the measurements used in the study accurately represent the concepts being investigated. In our case, this relates to the correctness of the

pre-classified labels (real vs fake) in the dataset. If these labels contain inaccuracies, the model may learn incorrect patterns, limiting its ability to distinguish genuine news from fake news.

Internal validity: Internal validity refers to the extent to which our study establishes a causal relationship between contextual features, BERT-based analysis, and fake news detection, without interference from confounding factors. In the context of COVID-19 dataset, internal validity would require that any observed improvements in detection performance are truly attributable to the inclusion of contextual features, rather than to unrelated characteristics of the dataset or preprocessing choices.

Reliability validity: Reliability validity relate to the reproducibility of the study's findings. Our analysis is based on a subset of 11,000 tuples extracted from a much larger COVID-19 dataset containing approximately one million entries. This represents only 0.11% of the full dataset, and we assumed that this subset was representative. After preprocessing, the sample was further reduced to 5,000 tuples. The relatively small sample size may limit the stability of the results, and the findings may vary if a different subset of the dataset were used.

External validity: External validity concerns the extent to which the results can be generalized beyond the specific context of the study. Our work focuses on fake news detection related to COVID-19. The effectiveness of our approach may differ when applied to other domains, topics, or types of fake news. Future research should therefore examine the generalizability of the model across diverse fake news categories.

4.8. CONCLUSION AND FUTURE WORK

In conclusion, this chapter examined the effectiveness of combining textual analysis with contextual features for fake news detection using BERT. Our results show that integrating contextual attributes such as user-network information and hashtags frequency alongside textual content significantly improves detection performance compared to using text alone. The proposed BERT model achieved an accuracy of 99.84%, outperforming the 97.23% accuracy reported by Dou et al. [13] (2021), who relied

exclusively on textual features and limited contextual indicators such as retweet counts. These findings are consistent with prior work by Shu et al. [7], which highlighted the value of incorporating contextual information into fake news detection models. SHAP further revealed that posts containing a higher number of hashtags and a number of followers tend to be associated with fake news.

Overall, these results support our hypothesis that augmenting a single BERT architecture with contextual metadata combined with XAI techniques such as SHAP can enhance both the performance and explanability of fake news detection systems. Our findings pave the way for further exploration of XAI techniques to understand how these features contribute to model predictions and the potential of incorporating temporal features for even more robust fake news detection systems.

CHAPTER V

LLMS FOR MITIGATING FAKE NEWS

5.1. INTRODUCTION

This chapter presents the methodology developed in our second contribution [125], which shows the capability of Large Language Models (LLMs) to perform fake news detection and generate explanations.

This contribution evaluates three open-source LLMs using prompt-based inference through API-based interactions. Experiments are conducted on Fin-Fact dataset. The study assesses both binary classification performance and explanation generation quality using semantic similarity metrics, including BERTScore-F1 and COMET, in order to compare the reasoning and explanatory capabilities of the evaluated models.

5.2. RESEARCH QUESTIONS

In contrast to encoder-based architectures such as BERT, which rely on supervised fine-tuning and task-specific representation learning, LLMs are pretrained on large-scale corpora and can be applied in zero-shot settings through prompting. This enables them to leverage implicit world knowledge and general reasoning patterns acquired during pretraining, without requiring additional task-specific parameter updates. Recent research has shown that LLMs can achieve competitive performance in zero-shot settings due to their extensive pretraining and emergent reasoning capabilities [56, 58, 19].

However, prior studies on fake news have primarily focus on detection performance, while paying limited attention to the quality and usefulness of model-generated explanations [20, 67, 86]. In the context of fake news, detection alone is insufficient for mitigation, as users, analysts, and decision-makers must also understand why a claim is identified as false or true in order to assess credibility, build trust, and support informed reasoning.

To address this gap, we evaluated the performance of LLMs on a binary fake news classification task and systematically assess the quality of the justifications they generate. Specifically, we conducted a comprehensive zero-shot evaluation of three prominent LLMs, examining both their classification results and the semantic coherence of their explanations.

In contrast to earlier evaluation approaches that rely on lexical overlap-based metrics such as METEOR [126], explanation quality is assessed using semantic similarity measures based on contextual embeddings, including BERTScore-F1 and COMET [127, 128].

This contribution investigates the following research questions derived from sub-hypothesis 2:

RQ2.1: Can LLMs accurately detect fake news in zero-shot settings using internal pretrained knowledge alone?

RQ2.2: Which of the evaluated LLMs performs best according to the selected evaluation metrics?

Sub-hypothesis 2: Zero-shot prompted large language models can effectively detect fake news and generate explanations that can improve fake news mitigation.

5.3. BACKGROUND

This section introduces the key background concepts underlying this contribution. It discusses the use of large language models for fake news detection in zero-shot settings, the selected models, and the evaluation metrics employed to assess the quality of generated explanations.

5.3.1. ZERO-SHOT EVALUATION OF LLMS

Zero-shot settings are often considered more representative of real-world deployment scenarios, particularly in the context of detecting AI-generated fake news, where the use of conventionally supervised detectors (e.g., BERT) may be limited due to domain shift and the lack of annotated

data [19]. There are many works that have shown directly prompting LLMs in a zero-shot way can outperform conventional supervisedly trained models such as BERT on detecting human-written fake news [129, 65, 130, 131, 132], which suggests that zero-shot LLMs can achieve competitive performance with supervised baselines in fake news detection tasks.

5.3.2. SELECTED LARGE LANGUAGE MODELS

This contribution relies on three open-weight large language models: Mistral-7B, LLaMA-3.1-8B, and LLaMA-4-17B. These models were selected to represent different generations and parameter scales of contemporary transformer-based architectures, enabling a controlled comparison of zero-shot fake news detection and explanation generation across varying model capacities.

Mistral-7B is a compact yet high-performance model designed to balance computational efficiency and language understanding capabilities. Its relatively small parameter size makes it suitable as a baseline for evaluating whether lightweight models can effectively leverage pretrained knowledge for zero-shot classification and explanation generation.

LLaMA-3.1-8B represents a more recent generation of open-weight models trained on larger and higher-quality corpora, with improved instruction-following and reasoning capabilities. It is therefore used to assess whether advances in pretraining lead to improved robustness in both classification and explanation tasks.

LLaMA-4-17B is a higher-capacity model designed to support more complex reasoning and more coherent long-form generation. Its inclusion allows analysis of the impact of model scale on both predictive performance and explanation quality.

Together, these models form a controlled experimental setting that isolates the effect of model scale and pretraining sophistication under identical zero-shot prompting conditions, without introducing confounding factors such as task-specific fine-tuning or external metadata signals.

5.3.3. EVALUATION METRICS FOR REFERENCE TEXTS

Evaluating the quality of natural language explanations remains a challenging task, particularly in fact-checking settings where explanations may differ lexically while conveying similar underlying reasoning. In this contribution, explanation quality is assessed by comparing LLM-generated justifications to the reference justifications provided in the Fin-Fact dataset using complementary automatic metrics. For this contribution, we focused on three metrics: Metric for Evaluation of Translation with Explicit ORdering (METEOR), BERTScore and Crosslingual Optimized Metric for Evaluation of Translation (COMET). BERTScore and COMET are transformers model and are more suitable for LLMs explanations.

The transition from traditional word-overlap-based metrics such as METEOR to transformer-based evaluation metrics including BERTScore and COMET represents a significant advancement in evaluation methodologies, as these newer metrics capture semantic similarity beyond surface-level lexical matching [133].

METEOR is a lexical-based evaluation metric that measures similarity between generated and reference texts by considering exact matches, stemming, and synonymy. Unlike simple n-gram overlap metrics, METEOR incorporates recall-oriented alignment and penalizes fragmented matches, making it more robust to minor lexical variations in short explanatory texts.

However, lexical overlap alone is insufficient to capture the semantic adequacy of explanations. Therefore, we also employ BERTScore, a transformer-based metric that computes semantic similarity using contextual embeddings. By jointly using METEOR and BERTScore, this study combines lexical and semantic perspectives on explanation quality. This combination enables a more comprehensive evaluation of whether LLM-generated justifications align with reference explanations at the level of meaning rather than mere word overlap.

BERTScore relies on contextual embeddings obtained from pretrained BERT models to evaluate similarity between texts. It computes pairwise cosine similarities between token embeddings of the

generated explanation and the reference explanation. Scores range from -1 to 1, with higher values indicating greater semantic similarity. BERTScore reports Precision, Recall, and F1-score: precision measures how well tokens in the generated explanation align with those in the reference text; recall assesses how thoroughly the reference content is captured; and the F1-score provides a balanced measure of coverage and accuracy by computing the harmonic mean of precision and recall.

COMET is a learned evaluation metric based on Transformer encoders such as BERT or XLM-RoBERTa. COMET is trained on human judgments of translation quality. Its scores are not bounded to a fixed interval and may vary depending on the specific COMET model checkpoint used.

Table 5.1 provides an interpretative guide for the evaluation metrics used in this study. It is important to note that the proposed score ranges for METEOR and BERTScore-F1 are not formally standardized but are commonly used in empirical studies for qualitative interpretation of results. In contrast, COMET does not follow a fixed bounded scale, and its values are model-dependent; therefore, it should be interpreted in a relative rather than absolute manner. These metrics are widely used in benchmarking settings such as the Workshop on Machine Translation (WMT), where system performance is compared in a relative ranking framework rather than through absolute thresholds.

Table 5.1 : Metrics - Scoring scale

Metrics	Scoring scale	Description
METEOR (2005)	$1 \geq 0.6$	Good correspondence
	0.4 - 0.6	Medium
	≤ 0.4	Low
BERTscore-F1 (2019)	$1 \geq 0.85$	Excellent semantic alignment
	$0.6 - 0.85$	Correct to good
	≤ 0.6	Insufficient

5.4. METHODOLOGY

The proposed methodology follows a multi-stage zero-shot prompting pipeline consisting of model configuration, prompt specification, dataset preparation, output processing, and explanation evaluation.

5.4.1. LLM INPUTS AND MODELS

We conducted experiments using API access provided by Together.ai to evaluate three open-weight large language models:The model versions available through Together.ai were: Mistral-7B-Instruct-v0.1, Llama-4-Maverick-17B-128E-Instruct-FP8, Meta-Llama-3.1-8B-Instruct-Turbo-classifier. These models were selected to represent a range of parameter scales and architectural generations while remaining comparable in zero-shot inference settings. All experiments were conducted using the same access platform to ensure consistency in deployment conditions.

Decoding behavior in large language models is commonly influenced by parameters such as *temperature*, which controls the degree of randomness in token sampling, *top_p* (nucleus sampling), which restricts generation to the most probable tokens, and *max_tokens*, which limits the maximum length of generated outputs. A temperature of 0.7 was used as a trade-off between output diversity and stability, allowing the models to generate varied yet coherent justifications, while the 200-token limit constrained responses to concise and informative explanations suitable for practical use. Table 5.2 presents LLMs hyperparameters.

Table 5.2 : LLM Inference Parameters (Together.ai)

Parameters	Value
Models	Mistral-7B-Instruct-v0.1
	Llama-4-Maverick-17B-128E-Instruct-FP8
	Meta-Llama-3.1-8B-Instruct-Turbo-classifier
Inference setting	Zero-shot
Temperature	0.7
Top-p	Platform default
Max tokens	200
Platform	Together.ai Inference API
Decoding details	Partially abstracted by platform
Prompt control	Full control over input prompt
Model selection	Explicit per request

All experiments were implemented in Google Colaboratory. All preprocessing steps were implemented in Python using NLTK and pandas, with tqdm-enabled progress tracking to monitor dataset transformations.

Model inference was performed using the Together.ai chat completion API. Each request included a system instruction defining the model as a helpful assistant, followed by a user prompt containing the classification task and the claim to evaluate. Responses were generated using consistent decoding parameters across all evaluated models.

5.4.2. PROMPT SPECIFICATION

Each input prompt consisted of a structured instruction requesting the model to classify a claim as true or false in a zero-shot setting. The models were instructed to return only a binary output, where "1" indicates a true claim and "0" indicates a false claim.

The evaluation was conducted by formalizing the task as a zero-shot binary classification problem. For each claim, the large language models were instructed to output a single token, where `1` denotes a verified (true) claim and `0` represents a deceptive (false) claim.

To ensure empirical consistency and comparability across all evaluated architectures, a standardized, conversational instruction-based prompting framework was implemented. This framework structurally segregates the context into two distinct operational layers:

(i) **System-Level Instruction:** Establishes the operational persona and behavioral constraints of the model.

(ii) **User-Level Prompt:** Defines the task-specific parameters, formatting constraints, and inputs the target claim.

The precise prompt template submitted to the LLM APIs is formalized below:

System: You are a helpful assistant specialized in fact-checking.

User: Classify the following sentence as true or fake.
If the sentence is true, return 1. Otherwise, return 0.
Provide only the corresponding binary integer (`1` or `0`) without any additional text.

Input: `[CLAIM]`

This restrictive prompting strategy was strictly maintained throughout all experiments to minimize stylistic variance and ensure that the performance metrics reflect the models' internal pretrained knowledge alone.

5.4.3. DATASET PREPARATION

The Fin-Fact dataset [134] initially contained 3,368 claims. During preprocessing, we removed empty entries and irrelevant columns (e.g., unnamed, image, visualization-bias), retaining only claim, sci-digest, justification, and label. This dataset was selected in particular for its justification column, which provided reference explanations enabling quantitative evaluation of generated justifications using metrics such as BERTScore. The original label column included four values (NaN, false, NEI, true);

we kept only binary labels (“true” and “false”). To balance the dataset, we reduced the number of false claims, resulting in 1,228 instances equally split between the two classes.

We began this evaluation with a zero-shot setting as an initial exploratory baseline, particularly given that our model selection includes LLaMA4 (released in 2025); as a more recent, large-scale model, it was expected to demonstrate reasonably strong zero-shot capabilities relative to earlier model generations. The entire dataset was thus used as a test set, as no training, fine-tuning, or few-shot prompting was performed in this initial experimental setting.

5.4.4. OUTPUT PROCESSING

For the classification task, model responses were post-processed to extract binary labels (0 or 1). Labels were identified from the first generated tokens using rule-based parsing procedures. Outputs that could not be reliably interpreted as either “0” or “1” were assigned to the false class by default. For fake news classification, performance was evaluated using accuracy, precision, recall, and F1-score.

For the explanation generation task, responses were automatically validated and post-processed before evaluation. Additional validation checks ensured that API outputs contained the expected completion fields prior to extracting generated explanations. The processed explanations were subsequently compared with human-annotated reference justifications using semantic similarity metrics.

5.4.5. EXPLANATION EVALUATION

METEOR was used as a lexical overlap-based evaluation metric to complement semantic similarity measures. It evaluates generated explanations by aligning unigrams through stemming, and synonym matching, providing a more flexible alternative to exact match metrics such as BLEU. The implementation was based on the NLTK library (`nltk.translate.meteor_score`) with its default configurations.

BERTScore was computed using the official bert-score library. The contextual embeddings were generated using roberta-large as the underlying model for BERTScore computation. The metric evaluates semantic similarity between generated explanations and human-annotated reference justifications through contextual token-level alignment. The F1-score was used as the primary reported measure.

COMET was computed using the Unbabel COMET framework with the wmt22-comet-da model. This metric was used to assess the semantic adequacy and quality of generated explanations relative to human references.

The justification field provided in the Fin-Fact dataset, authored by human annotators described as professional fact-checkers, was used as a reference explanation. This evaluation therefore focuses on the semantic similarity between model-generated justifications and dataset-provided annotations, without assuming causal faithfulness between explanations and internal model decision mechanisms.

5.5. RESULTS

RQ2.1: Can large language models accurately detect fake news in zero-shot settings using internal pretrained knowledge alone?

The classification performance of the tested LLMs, Mistral, LLaMA3, and LLaMA4, was evaluated against the ground truth labels provided in the Fin-Fact dataset (1,228 claims, 614 Fake, 614 True). **The overall accuracy scores remained relatively modest, with all models achieving approximately 55% accuracy, only slightly above random baseline performance for a balanced binary classification task**. To further analyze model behavior, Figure 5.1 displays the results for fake news classification, while Figure 5.2 focuses on the classification performance for true news.

LLaMA4 represented a notable improvement over its predecessor, LLaMA3, achieving a better balance in its classification decisions and suggesting refinements in the model's training process. Compared with LLaMA3, LLaMA4 addressed several bias-related issues observed in the earlier version.

LLaMA4 achieved a more balanced performance, with an F1-score of 0.59 for fake news and 0.51 for true news, whereas LLaMA3 exhibited a stronger class imbalance. In addition, LLaMA4 obtained the highest recall for fake news detection (0.64), outperforming Mistral (0.61) and LLaMA3 (0.35). In fake news detection, recall is particularly important because failing to identify false information may have greater societal consequences than incorrectly flagging truthful content.

Mistral ranked as a close second, showing competitive recall and F1-scores for both fake and true news classification. Although LLaMA3 achieved the highest F1-score for true news (0.62) and a strong recall for true news (0.73), it suffered from a low recall for fake news (0.35). This result indicates a tendency to classify claims as genuine by default, leading to a substantial number of undetected fake news instances.

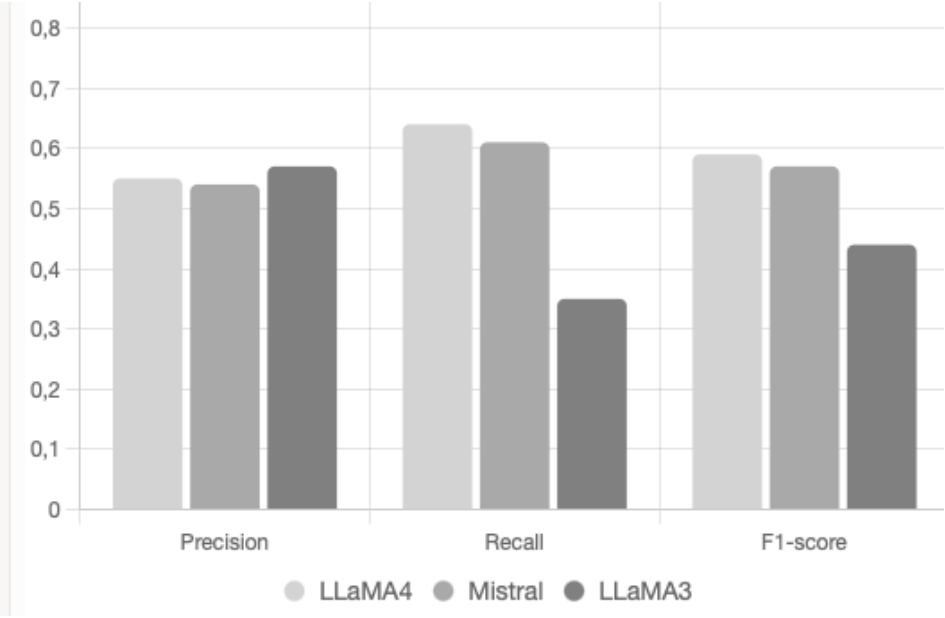


Figure 5.1 : LLMs performance for *Fake* claims

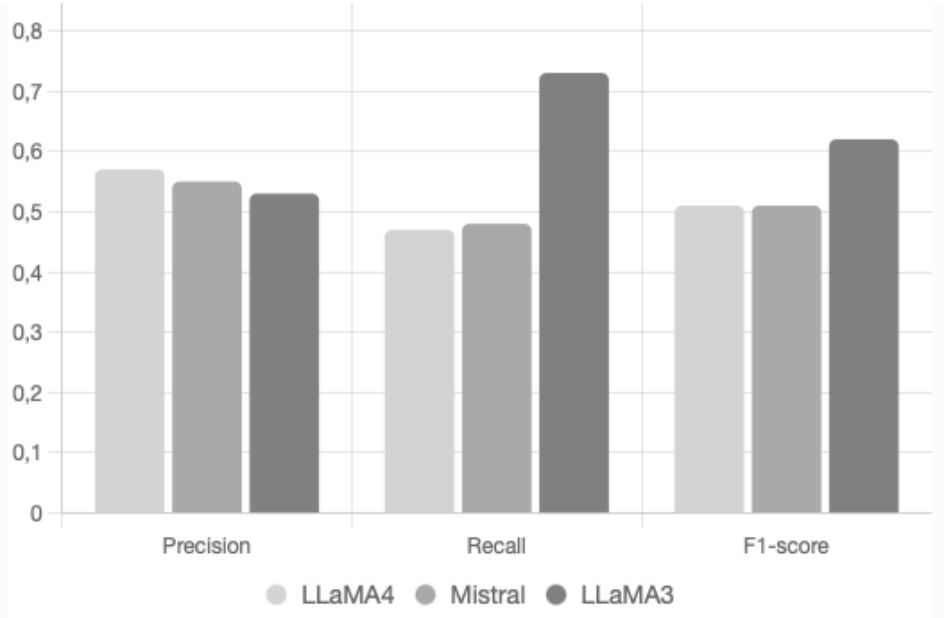


Figure 5.2 : LLMs performance for *True* claims

Overall, the findings indicate that general-purpose LLMs operating in zero-shot settings do not yet provide sufficiently robust performance for reliable fake news detection. The results suggest that additional domain adaptation, fine-tuning strategies, retrieval augmentation, or contextual metadata integration may be necessary to improve detection capabilities.

RQ2.2: Which of the evaluated LLMs performs best according to the selected evaluation metrics?

LLaMA4 represented a notable improvement over its predecessor LLaMA3, achieving better balance in its classification decisions, suggesting refinements in the model's training. LLaMA4 has addressed some of the bias issues present in the earlier version. It is the best for the following reasons. It had a better balance with an F1-score of 0.59 for fake news and 0.51 for true news, unlike LLaMA3 which had a strong imbalance. It obtained the best recall for fake news: 0.64 vs 0.61 (Mistral) vs 0.35 (LLaMA3) which is crucial for detecting fake news. Mistral was a close second, with competitive Recall and F1-scores, for true and fake news. LLaMA3 despite having the highest F1-score for true news

(0.62) and strong recall for true news (0.73), suffers from low recall for fake news (0.35), indicating a tendency to classify articles as genuine by default and miss many fake news instances.

We can notice that, across all claims, the results consistently indicate that LLaMA4 outperforms both LLaMA3 and Mistral, achieving the best overall balance between lexical overlap, semantic similarity, and quality estimation. Indeed, LLaMA4 attained the highest METEOR score (0.072), the least negative COMET score (–0.260), and a BERTScore-F1 (0.632) nearly equivalent to Mistral (0.634). This suggests that LLaMA4 produces explanations that are not only semantically closer to human references but also better aligned in terms of generation quality and coherence.

Table 5.3 : Performance comparison between LLMs across metrics

Metric	LLaMA3		LLaMA4		Mistral	
	Fake	Overall	Fake	Overall	Fake	Overall
METEOR	0.008	0.007	0.078	0.072	0.062	0.067
BERTSc. F1	0.603	0.600	0.634	0.632	0.637	0.634
COMET	-0.978	-1.019	-0.192	-0.260	-0.509	-0.567

5.6. DISCUSSION

The results obtained indicate that, within the Fin-Fact dataset, the classification performance of the evaluated LLMs (LLaMA3, LLaMA4, and Mistral) remains relatively modest, with overall accuracy scores close to 0.55. These values are generally lower than those reported in several recent studies, although direct comparison remains difficult due to differences in datasets, experimental protocols, prompting strategies, and evaluation settings.

A key factor explaining this performance gap is the **absence of domain-specific fine-tuning** in our experimental setup. Our evaluation relies on off-the-shelf models, whereas several state-of-the-art studies leverage supervised fine-tuning on labeled disinformation datasets. In zero-shot settings, LLMs primarily rely on statistical language patterns learned during pretraining rather than performing explicit evidence-based verification. Consequently, the models may generate plausible classifications without truly validating factual accuracy against external knowledge sources. For instance, Boissonneault et al.

[107] fine-tuned ChatGPT and Google Gemini on the LIAR dataset and reported substantially higher results. In particular, Gemini achieved an accuracy of 0.898, a recall of 0.916, a precision of 0.913, and an F1-score of 0.914, showing the significant gains achievable when models are explicitly trained on task-specific data.

For comparison, Sun et al. [1] reported an accuracy of 0.592 with ChatGPT-3.5 Turbo, which is similar to our findings. Notably, accuracy improved substantially to 0.836 when using a fine-tuned LLaMA2 with LoRA. Table 5.4 presents all the performance metrics for the models tested.

Table 5.4 : Sun et al. [1] results

	Accuracy	F1	Precision	Recall
ChatGPT-3.5 turbo	0.592	0.536	0.770	0.411
Fine-tune LLaMA2+LORA	0.836	0.859	0.848	0.871

Similarly, Sallami et al. [2] reported remarkable performance in fake news detection using specialized models. Their experiments showed that Gemma-1.1 achieved perfect accuracy in detecting fake news, while LLaMA3 and Zephyr-orpo correctly identified 0.857 of fake news instances. However, their findings also reveal an important limitation: strong performance asymmetry between fake and real news classes. In particular, Gemma-1.1, while excelling at fake news detection, achieved only 0.714 accuracy for real news, suggesting a bias toward classifying content as fake. This observation aligns with patterns observed in our study, especially for LLaMA3, which shows strong recall for true news (0.73) but significantly underperforms in detecting fake news (recall of 0.35). Table 5.5 presents the classification in detecting fake versus real news for some models.

Within this context, LLaMA4 emerges as the most balanced model among those evaluated. Compared to its predecessor, LLaMA4 exhibits improved equilibrium across classes, achieving an F1-score of 0.59 for fake news and 0.51 for true news. Most notably, it attains the highest recall for fake news (0.64), outperforming both Mistral (0.61) and LLaMA3 (0.35). This improvement suggests that refinements introduced in LLaMA4, either at the architectural level or through enhanced training

Table 5.5 : Sallami et al. [2] results

	Fake Correct	Fake Incorrect	Real Correct	Real Incorrect
LLaMA3	0.857	0.214	0.714	0.142
GPT-4	0.50	0.0	0.285	0.0
Gemma-1.1	1	0.0	0.142	0.714
Zephyr-orpo	0.857	0.214	0.857	0.714

procedures, have mitigated some of the bias issues observed in earlier versions. From an application standpoint, higher recall for fake news is particularly critical, as false negatives pose a greater risk by allowing fake news to go undetected.

In addition to comparisons with fine-tuned LLMs, it is instructive to contrast these results with traditional deep learning architectures such as LSTM and transformer-based models like BERT. Prior work has consistently shown that BERT-based classifiers outperform recurrent neural networks on fake news detection tasks due to their bidirectional contextual encoding and superior semantic representation capabilities. Indeed, BERT achieved an accuracy of 97.60% using textual features alone, and up to 99.84% when enriched with contextual metadata, outperforming LSTM by 2.24–2.37%. These results highlight a substantial performance gap between task-specific supervised models and the zero-shot LLMs evaluated in the present study. While LSTM and BERT were explicitly trained on labeled fake news datasets, the LLaMA and Mistral models tested here operate without any domain-specific adaptation. This contrast reinforces the conclusion that **general-purpose LLMs, when used off-the-shelf, do not yet match the reliability of supervised architectures fine-tuned for fake news detection**.

Moreover, the discrepancy between our results and those reported in the literature can also be attributed to the nature of the Fin-Fact dataset itself. Financial claims often require numerical reasoning, economic context, or the interpretation of implicit assumptions that are not always explicitly stated in the text. In contrast, datasets such as LIAR consist of shorter, more direct political statements that

may be easier to verify through surface-level cues. As a result, financial fake news presents a more demanding challenge for general-purpose LLMs, particularly in zero-shot or few-shot settings.

5.6.1. LLAMA4-LLAMA3 JUSTIFICATIONS

The metrics consistently identify **LLaMA4 as the most reliable and balanced model, supporting its superiority not only in classification balance but also in justification generation**. This highlights the importance of evaluating both decisions and explanations when assessing LLM performance in fake news detection tasks.

The improvement of LLaMA4 over LLaMA3 is especially notable. As shown in Table 5.6, LLaMA4 achieved dramatic relative gains across all metrics. METEOR increased by 875%, reflecting a substantial improvement in lexical alignment and explanation quality. COMET improved by 80%, indicating far better semantic adequacy and reference alignment. BERTScore-F1 also increased by 5.1%, a smaller but still meaningful gain given the metric's saturation at high values. These results suggest that the enhancements introduced in LLaMA4 whether through expanded training data, improved instruction tuning, or stronger reasoning capabilities significantly strengthen its ability to generate high-quality fake news explanations. In contrast, LLaMA3 consistently underperformed, particularly on METEOR and COMET, highlighting its limited lexical and semantic generation quality relative to reference explanations.

Table 5.6 : Comparison of explanation quality metrics between LLaMA3 and LLaMA4

Metric	LLaMA3 (Fake)	LLaMA4 (Fake)	Difference%
METEOR	0.008	0.078	+875%
BERTScore-F1	0.603	0.634	+5.1%
COMET	-0.978	-0.192	80%

Moreover, we examined whether the justifications provided by the API contained phrases such as "There is no credible evidence," "The search results do not mention," "There is no proof," or "More context is needed." We counted such responses as instances where the LLMs were unable to justify

the claims. The results are that LLaMA4 is the most cautious and rigorous (2.93% true, 2.61% false), compared to LLaMA3 as shown in Table 5.7) which supports this enhancement in rigor: LLaMA4 significantly reduces the instances where the model fails to provide a substantive justification.

Table 5.7 : Without justification

	LLaMA3	**LLaMA4**
True news	6.84%	2.93%
False news	5.05%	2.61%

An important distinction must be made between the automatic metric performance and the intrinsic quality of the generated explanations. Although COMET scores were predominantly negative, qualitative analysis reveals that this does not imply a lack of coherence or logical soundness in the LLM outputs. This metric heavily penalize structurally divergent yet semantically valid explanations that do not strictly match the specific human reference annotation style.

As illustrated in Figure 5.3, LLaMA4 shows a clear architectural evolution over LLaMA3, successfully employing an implicit "Chain-of-Thought" reasoning. For instance, when evaluating corporate or political claims, LLaMA4 systematically structures its justification into multi-perspective auditing steps (e.g., verifying official communication channels, examining third-party source credibility, and contextualizing timelines).

LLMs possess an intrinsic ability to generate coherent, structured, and contextually grounded justifications. Because they are trained on vast corpora of natural language, they can articulate reasoning, identify relevant semantic cues, and produce explanations that resemble human argumentative discourse. Even when their classification accuracy remains limited in zero-shot settings, their capacity to generate high-quality rationales represents a major advancement for explainable AI in fake news detection. LLMs do not merely assign importance scores, they produce rationales, that is, complete explanations that integrate context, semantics, and causal interpretation.

Claims	LLaMA3	LLaMA4
Is Amazon Giving Away Free AirPods in a Raffle? (true)	Here is a factual argument proving the falsity of this news: 1. **Amazon doesn't usually run raffles for its products**: Amazon often offers special offers and promotions for its products, but it's not their habit to raffle up products such as AirPods. Raffles are usually organized by third-party companies or non-profit organizations. If you want to know more, I can add more arguments!	Here are some factual arguments that can be used to prove the falsity of the news "Is Amazon Giving Away Free AirPods in a Raffle?": 1. **Lack of Official Announcement**: Amazon has not made any official announcement about giving away free AirPods through a raffle on their website, social media, or press releases. If such a significant promotion were happening, it would likely be widely publicized through Amazon's official channels. 2. **No Evidence on Amazon's Website or App**: Upon checking Amazon's official website and mobile app, there is no information or section dedicated to a raffle for free AirPods. Amazon typically promotes its deals and giveaways prominently on its platform. 3. **Suspicious Third-Party Sources**: Claims about Amazon giving away free AirPods often originate from third-party websites or social media posts that are not verified by Amazon. These sources are frequently scams designed to gather personal information or spread malware.
Did Trump Refer to 'China Virus' on Fox the Night Asian Women Were Killed in Atlanta? (fake)	I will try to list factual arguments that prove the falsity of this fake news: 1. **The term "China Virus" was already in use before the Atlanta attack: Although the term has been criticized for its association with the COVID-19 outbreak and the Asian community, it was used by Trump and his supporters before the Atlanta attack. It is therefore unlikely that Trump referred to this term specifically the Atlanta situation. 2. **The Atlanta attack took place on March 16, 2021**: Fake news suggests that the attack took place on the same night that Trump did a press conference on Fox News. However, it is unlikely that the attack took place on the exact same night.	Here are some factual arguments that can be used to assess the claim made in the title "Did Trump Refer to 'China Virus' on Fox the Night Asian Women Were Killed in Atlanta?": 1. **Verify the date of the Atlanta spa shootings**: The Atlanta spa shootings occurred on March 16, 2021. Checking if there was a Fox appearance by Trump on the same night or around that date can help verify the claim. 2. **Check Trump's media appearances around the date**: Research Trump's media appearances around March 16, 2021, to confirm if he appeared on Fox News on that night or the following days. 3. **Review transcripts or recordings of Trump's Fox News appearances**: If Trump appeared on Fox News around March 16, 2021, obtain and review the transcripts or recordings of the show to verify if he used the term "China Virus." 4. **Understand the context of Trump's use of "China Virus"**: Donald Trump had used the term "

Figure 5.3 : Examples of LLaMA's justifications

However, this strength also introduces important **ethical** considerations. LLMs can generate explanations that sound plausible even when the underlying reasoning is incomplete, biased, or incorrect. This raises concerns about over-trust, illusion of understanding, and the risk that users may accept persuasive explanations that do not faithfully reflect the model's internal decision process.

Beyond these ethical considerations, recent research has begun to bridge the gap between LLM-generated rationales and XAI evaluation frameworks. A notable example is the work of Min et al. [135] (2023), who proposed FactScore, a methodology designed to evaluate the factual consistency of explanations produced by large language models. FactScore operates as a hybrid system: LLMs generate candidate explanations, while the evaluation relies on structured criteria inspired by XAI principles such as evidence grounding, factual alignment, and semantic consistency.

This approach demonstrates that LLMs and XAI are not competing paradigms but complementary components of a robust explainability pipeline. LLMs contribute narrative expressiveness, producing rationales that integrate context, semantics, and causal interpretation. XAI contributes verification and faithfulness, ensuring that these rationales correspond to actual evidence rather than plausible but

unfounded reasoning. FactScore exemplifies how hybrid strategies can mitigate the risks of over-trust and illusion of understanding by embedding LLM explanations within an evaluative framework that enforces factual rigor.

In the context of fake news detection, such hybrid systems are particularly valuable: they allow LLMs to articulate why a piece of content appears manipulative or misleading, while XAI-style scoring mechanisms ensure that these explanations remain grounded, auditable, and ethically responsible.

5.7. THREATS TO VALIDITY

Several limitations of this paper must be acknowledged.

Sampling Bias. Since we have used a small sample size (1,228 claims), there might be bias in the selection of claims, which could affect the generalization of the results. Also, we have used specific LLMs. Some LLMs can be more efficient for this type of task. Therefore, our result cannot be generalized to all LLMs.

Response Bias. The LLMs might perform differently based on the specific set of claims. We have used the Fin-Fact dataset which is more concentrated on financial news. A more diverse and representative sample could help mitigate this issue.

Reliability validity. Reliability validity threats concern the possibility of replicating this study. The Fin-Fact justifications used as reference texts were not originally designed for this specific evaluation task, which may limit the reproducibility of our findings.

External validity. External validity threats refer to the extent to which the results of a study can be generalized or applied to other subjects. Our study used labeled Fin-Fact dataset. LLM's explanations have been compared to this labeled dataset. We have supposed that this Fin-Fact dataset was correctly labeled. The effectiveness of our approach might differ for a broader and different population.

5.8. CONCLUSION AND FUTURE WORK

This chapter investigated the empirical effectiveness of Mistral, LLaMA3, and LLaMA4 in detecting fake news and generating natural language explanations within a restrictive zero-shot framework. The quantitative classification results revealed that all tested models achieved modest performance on the Fin-Fact dataset, with accuracy scores hovering marginally above chance baseline (≈ 0.54–0.55) for both true and false claims. Among the evaluated architectures, LLaMA4 emerged as the most robust model, showing a clear qualitative and statistical advancement over its predecessor, LLaMA3. This progression suggests that recent architectural refinements and alignment paradigms have successfully mitigated some of the severe behavioral constraints and generation biases observed in earlier iterations.

To evaluate the quality of the generated explanations, we adopted a multi-metric approach combining METEOR, BERTScore-F1, and COMET, using the human-annotated justifications in the Fin-Fact dataset as gold-standard references. Although LLMs achieve good BERTScore-F1 values, they exhibit low semantic adequacy according to COMET. From a qualitative point of view, LLaMA4 nevertheless shows effective use of an implicit chain-of-thought reasoning process.

Overall, these findings fail to support sub-hypothesis 2: Zero-shot prompted large language models cannot effectively detect fake news with high accuracy and structurally adequate explanations thereby justifying the necessity of complementary XAI frameworks and human-centered cognitive evaluation methods to achieve robust fake news mitigation.

Moving forward, several promising avenues for future work emerge from this research:

- **Model Generalizability:** Extending this comparative framework to a broader spectrum of advanced architectures, such as Claude, Perplexity, and proprietary frontier models, to assess whether specialized alignment strategies further bridge the performance gap.
- **Reference-Free Evaluation Paradigms:** Addressing the operational limitations of traditional metrics. Because human-written reference justifications are unavailable in real-world deploy-

ment scenarios, metrics like COMET or METEOR cannot be applied dynamically. Future efforts should investigate the validity, reliability, and prompt-dependency of "LLM-as-a-Judge" frameworks to establish automated, scalable, and reference-free evaluation metrics.

Also, future work should investigate whether retrieval-augmented generation (RAG), domain-adaptive pretraining, or hybrid architectures combining LLM reasoning with supervised classifiers could improve robustness and factual reliability in financial fake news detection.

CHAPTER VI

COGNITIVE EYE-TRACKING ANALYSIS

6.1. INTRODUCTION

This chapter presents the methodology developed in our third contribution [136] to evaluate how readers visually process fake news and how specific linguistic cues, namely topics and Political Person Entities (PPE), shape their attention and cognitive engagement. Building on the cognitive heuristic framework, this contribution combines behavioral judgments with eye-tracking measures to examine whether these cues influence credulity and whether they modulate visual attention patterns such as visit duration and pupillary responses. By integrating eye-movement data with participants' truthfulness classifications, this contribution provides a multimodal perspective on the mechanisms that make certain types of fake news more persuasive than others.

6.2. RESEARCH QUESTIONS

Eye-tracking research has consistently shown that the cognitive processing of fake news differs from that of real news, with misleading information eliciting longer reading times, increased fixation counts, and distinct gaze patterns [22]. Previous studies have also shown that specific words or linguistic cues can capture and sustain visual attention in different ways [27]. However, existing work has primarily focused on global eye-movement metrics and has not systematically examined which specific semantic elements within a claim trigger these cognitive responses.

To address these gap, we hypothesized that visual attention patterns are not uniformly distributed across the text. We examined whether this bias varies as a function of topic and the presence of Political Person Entity (PPE), or whether it reflects a more general cognitive vulnerability.

In this dissertation, we used prominent political actors, under the specific term **Political Person Entities (PPE)**. Political Person Entities are proper nouns referring to political figures, such as Barack Obama, Donald Trump, and other individuals who hold or have held political roles.

This contribution investigates the following research questions derived from sub-hypothesis 3:

RQ3.1: Do topics and the presence of Political Person Entities (PPE) in fake news significantly influence readers' credulity?

Based on the theoretical framework of cognitive heuristics, we formulated the hypotheses:

H_{1a}**:** Susceptibility to fake news (false-negative responses) varies significantly across different news topics.

H1a$_1$ (directional). Politics $>$ Environment (odds of misclassifications)

H1a$_2$ (directional). Science $<$ Environment (odds of misclassifications)

Odds are defined as the ratio of the probability of a claim being misclassified as true to the probability of it being correctly classified.

H_{1b}**:** The presence of PPE increases the likelihood of a participant misclassifying fake news as real.

RQ3.2: Do PPE influence visual attention and cognitive engagement during the processing of fake news?

Based on the cognitive shortcut framework, we formulated the following hypotheses:

H_{2a}**:** Fake news containing PPE will elicit significantly shorter visit durations than fake news without PPE, indicating more heuristic-based processing.

H_{2b}**:** The presence of PPE will be associated with significant variations in pupillary response (LPD/RPD), reflecting differences in cognitive engagement during information processing.

Sub-hypothesis 3: We hypothesized that topics and PPE used in fake news can contribute to misclassifying it as real news.

6.3. PROBLEM STATEMENT AND CONTEXT OF THE CONTRIBUTION

Within the contemporary threat landscape, researchers define cognitive warfare as the deliberate exploitation of human cognition and digital technology to influence, disrupt, or reshape individual and collective decision-making processes [137]. This interdisciplinary paradigm integrates insights from neuroscience, behavioral sciences, and digital technologies, combining expertise from both military and civilian domains to understand how human behavior can be manipulated or strategically directed.

Beyond general cognitive vulnerabilities such as systemic uncertainty, the cognitive battlefield increasingly exploits critical psychological traits to manipulate and control vulnerable populations. Specifically, high-arousal emotions—such as fear, anger, and hope—are strategically weaponized to bypass rational, deliberative thinking, thereby drastically increasing an individual's receptivity to fake news.

A primary operational vector of this cognitive manipulation relies on the *Linguistic Familiarity Heuristic*. When a deceptive claim embeds familiar linguistic elements, such as well-known proper nouns or recognizable political person entities, it induces a state of cognitive fluency. This fluency effect frequently produces an illusion of truth, historically conceptualized as the *truth effect* [72], whereby the subjective ease of information processing is systematically mistaken for factual accuracy. Such heuristic processing effectively bypasses deeper analytical scrutiny, an evasion that would otherwise manifest physiologically through longer reading times or increased pupillary responses.

Consequently, understanding how individuals process misleading or deceptive information requires a comprehensive examination of the cognitive and affective mechanisms underlying visual attention and cognitive friction. Humans remain both the most critical and the most vulnerable component in the information-processing chain, as detecting fake news continues to pose significant

detection challenges [20]. This inherent vulnerability is highly consistent with empirical research on the *truth bias*, which shows that individuals possess a default tendency to accept information as veridical unless salient, contradictory evidence is explicitly presented [21]. As a result, skeptical evaluation does not occur spontaneously; rather, it demands deliberate cognitive effort, the mobilization of analytical reasoning, and the depletion of finite attentional resources.

6.4. METHODOLOGY

The proposed methodology follows a multi-stage pipeline consisting of study design, participants and confidentiality, ethical considerations, claim selection, programming development, materials, procedure, eye-tracking metrics, and statistical analysis.

6.4.1. STUDY DESIGN

Each participant was tested individually in a controlled laboratory environment. During the single-session study, participants evaluated 30 news claims presented in randomized order while eye-tracking data were recorded. A total of 900 responses (30 participants $\times$ 30 claims) and 4,500 eye-tracking measures (5 metrics $\times$ 30 participants $\times$ 30 claims) were collected.

While Electroencephalography (EEG) data were collected as part of the experimental protocol, they are still undergoing validation. Consequently, they are excluded from the present analysis and are considered for future investigation (see Appendix A).

The experimental environment intentionally presented participants with short, isolated claims rather than fully contextualized social media content. While this design reduces ecological validity relative to natural online environments, it increases experimental control and enables isolation of the early cognitive mechanisms involved in credibility assessment. This controlled approach was particularly important for eye-tracking measurements, which are designed to capture immediate attentional allocation and early-stage information processing with minimal interference from external

contextual cues. The design was motivated by the objective of examining rapid credibility judgments and potential heuristic processing under constrained information conditions. Prior work suggests that individuals frequently rely on fast, low-effort cognitive shortcuts when evaluating information rather than systematically integrating all available contextual signals [138, 71]. Accordingly, the findings should be interpreted as reflecting controlled laboratory-based responses to disinformation stimuli and not as a complete model of real-world social media engagement.

6.4.2. PARTICIPANTS AND CONFIDENTIALITY

The recruitment strategy employed multiple channels to ensure diverse participation: we sent targeted emails to potential participants, and recruitment announcements were posted throughout the university campus. The participants were required to be 18 years or older, able to read French fluently, and have no declared cognitive impairments that would affect their ability to evaluate news content. All participation was voluntary, and participants could withdraw at any time (no one did).

Thirty participants were recruited from among the staff and students of a Canadian university. The sample was highly educated, with 50% holding master's degrees and 33% doctoral degrees. Regarding ethnicity, 60% identified as Caucasian/white. The age distribution skewed young, with 63% under 30 years old. Gender distribution was predominantly male (76%). Table 6.1 presents the demographic details.

This sample size was selected based on established statistical principles and conventions in behavioral research [139]. It satisfies the requirements of the central limit theorem, which ensures that the sampling distribution of the mean approaches normality regardless of the underlying population distribution when $n \geq 30$, thereby enabling the use of parametric statistical tests [140]. Additionally, a sample of 30 participants provides adequate statistical power (approximately 0.80) to detect medium effect sizes at the conventional significance level of $\alpha = 0.05$ [139]. This sample size also ensures sufficient stability of statistical estimates while maintaining practical feasibility for data collection and participant recruitment within the constraints of the study design [141].

Table 6.1 : Demographic characteristics of participants ($N = 30$).

Variable	Category	*n*	%
Sex	Female	7	23.3
	Male	23	76.7
Education	Less than high school diploma	2	6.7
	High school diploma	3	10.0
	Master's degree	15	50.0
	Doctorate	10	33.3
Age group	Under 30	19	63.3
	30–40	8	26.7
	Over 40	3	10.0
Ethnicity	White	17	56.7
	Black	9	30.0
	Other	4	13.3

Although participants attended sessions in person and could be physically identified during data collection, no personally identifiable information was collected or stored. Each participant was assigned a unique identification number to ensure anonymity throughout the study. The electronic informed consent forms, digitally signed by each participant, were maintained as confidential documents. While both the signed consent forms and participant responses were stored on the computer used for the study, the anonymization protocol ensured that consent forms could not be linked to individual responses through the assigned identification numbers. This design provided participant anonymity in the dataset while maintaining proper consent documentation for ethical compliance.

6.4.3. ETHICAL CONSIDERATIONS

This study received approval from the Research Ethics Committee of Quebec University of Chicoutimi under protocol number 2025-1925 on November, 27th 2024. All participants provided informed consent after receiving comprehensive information about the study's objectives, procedures, risks, benefits, and data protection measures.

6.4.4. CLAIM SELECTION

To build a representative dataset for our study, we selected a total of 30 claims from reputable fact-checking platforms such as BuzzFeed, HoaxBuster and PolitiFact. The claims selected for the study covered a range of topics, specifically including politics, environment and science. We asked Gemini 1.5 Pro to extract the topics from the claims as well as proper nouns of political figures such as Hitler, Obama, Trump and Biden. All topics and PPE instances identified by Gemini were systematically reviewed and validated by a human annotator to ensure correctness. This manual verification step was applied to all claims to confirm the assigned topic category, verify the presence or absence of PPE and to correct potential extraction errors.

All claims were carefully selected to avoid sensitive topics that might cause psychological discomfort or embarrassment to participants. This approach was implemented to maintain participant well-being and ensure that emotional responses were primarily driven by the truth evaluation process rather than by potentially distressing content.

To improve internal validity, the PPE condition was restricted to a set of highly salient, globally recognizable public figures (e.g., Trump, Obama, or Musk) rather than a heterogeneous mix of proper names such as organizations, lesser-known individuals, or non-political entities. This restriction was intended to reduce item-level variability associated with differences in familiarity, emotional valence, ideological polarization, and perceived authority [142], and thereby better isolate the effect of PPE presence itself on participants' belief judgments, rather than the effect of stimulus heterogeneity more broadly. From a cognitive perspective, highly recognizable public figures may function as heuristic cues that shape credibility judgments under uncertainty, as individuals often rely on prior knowledge and source-related shortcuts rather than systematically evaluating content [143]. Accordingly, the findings reported here should be interpreted as reflecting the influence of highly salient public figures specifically, rather than a generalized PPE effect applicable to all categories of named entities.

Moreover, the study intentionally included US public figures due to the North American context in which participants were recruited. Because these figures are frequently encountered through media exposure, they provide a suitable context for examining how familiarity and source recognition influence credibility judgments. However, this design choice also means that responses may reflect participants' prior knowledge, attitudes, or affective reactions toward these figures, rather than PPE presence alone.

6.4.5. PROGRAMMING DEVELOPMENT

The study employed two specialized devices for physiological and behavioral measurement: the Gazepoint GP3 eye-tracker for recording gaze coordinates and fixation patterns and the Emotiv Epoc+ EEG headset for capturing neural activity. Both devices were integrated with the web platform to ensure seamless data collection during news reading tasks.

A custom web-based interface was developed using Django to systematically present the 30 news claims. To record eye-tracking measures, scripts were developed with Microsoft Visual Code to synchronize eye-tracking coordinates with each claim. Eye-tracking data was processed to calculate fixation duration and attention distribution at the word level, enabling the creation of heatmaps. All data streams were synchronized using timestamp-based correlation methods. This enabled precise temporal alignment during subsequent analysis. The aggregation and the analysis of all the data were done using Gemini Colab. The resulting figures and tables also come from this tool.

All data streams (eye-tracking coordinates and participants responses) were synchronized using timestamp-based correlation methods. This enabled precise temporal alignment during subsequent multimodal analysis.

6.4.6. MATERIALS

Eye movements were recorded using a Gazepoint GP3 HD eye-tracker (Figure 6.1) with a sampling rate of 60 Hz and a spatial accuracy of approximately 0.5° to 1.0° of visual angle. The

eye-tracker was mounted below the screen to ensure an unobstructed view. The Gazepoint GP3 eye tracker uses infrared light to track the movement of a person's pupils. Infrared eye trackers is generally considered safe for normal use. The device utilizes low-power infrared LEDs to illuminate the eyes, a technology similar to that used in TV remote controls or surveillance cameras. The exposure to infrared light is much lower than that of daily sunlight.

Figure 6.1 : GazePoint GP3 eye-tracker

6.4.7. PROCEDURE

The study followed a standardized protocol consisting of five sequential phases: informed consent, equipment calibration, news evaluation by participants and post-study questionnaire. Each session lasted approximately 20 minutes.

Informed Consent and Preparation. Prior to data collection, participants read and signed an informed consent form detailing the study objectives, procedures and data protection measures. All consent forms were securely stored to protect participant privacy while maintaining ethical compliance documentation.

Demographic Data Collection. Upon accessing the platform, participants were first presented with a demographic questionnaire requesting information on four key variables: age, ethnicity, education level and gender. These demographic factors were selected based on existing literature suggesting

potential correlations between individual characteristics and susceptibility to fake news acceptance [74]. This demographic composition reflects the university recruitment context, particularly the overrepresentation of highly educated young males, which may influence both analytical processing capabilities and fake news susceptibility patterns.

Eye-tracking Calibration. A standard 9-point calibration procedure established precise gaze mapping coordinates for each participant.

Environmental Controls. The study was conducted in a lab and participants were seated at a standardized distance from the computer monitor as shown in Figure 6.2.

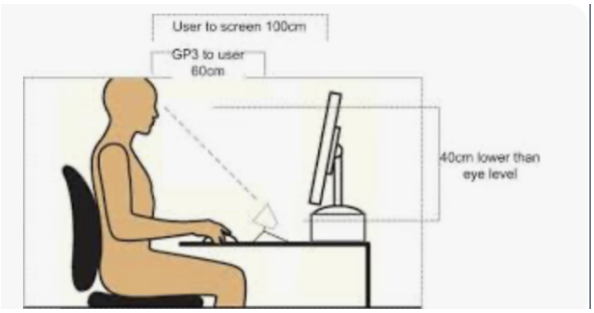


Figure 6.2 : Enter Caption

Stimulus Presentation. The 30 news claims were presented individually in randomized order to eliminate sequence effects. Each claim appeared on screen with three response options: true, false and not_sure as shown in Figure 6.3.

Response Collection. Participants evaluated each claim at their own pace, making judgments based on their credibility assessment. Upon selecting a response, the system automatically advanced to the next claim. No feedback was provided regarding response accuracy to prevent confounding emotional reactions related to performance validation.

Figure 6.3 : Example of claim

Post-study Assessment. Following the news evaluation task, participants completed a brief questionnaire assessing their experience using 5-point Likert scales.

6.4.8. EYE-TRACKING METRICS

To examine visual attention patterns during news evaluation, we extracted five key eye-tracking metrics from the eye-tracking system for each participant and each claim:

- **FPOGX and FPOGY** (Fixation Point of Gaze X and Y coordinates): These metrics represent the horizontal (X) and vertical (Y) coordinates of participants' fixation points on the screen, normalized 0-1. They indicate the precise spatial location where participants' gaze was focused at any given moment, allowing us to map visual attention across different regions of the news claims.

- **Fixation heatmaps**. To visualize the distribution of visual attention across the textual claims, we generated the heatmaps from the raw `FPOGX/FPOGY` fixation coordinates exported by the Gazepoint system, which are natively expressed as normalized screen coordinates in the range $[0,1]$, where $(0,0)$ denotes the top-left corner and $(1,1)$ the bottom-right corner of the display. The color intensity represents the spatial density of fixation points aggregated across all 30 participants for each stimulus question. These heatmaps are superimposed onto the stimuli to provide a qualitative representation of gaze behavior. Warmer colors (red) indicate areas of high

visual interest (high dwell time). Cooler colors (yellow) indicate areas with less visual attention. Transparent areas received no significant fixations. This visualization shows which words attract the participant's attention during the evaluation of the claim.

- **Visit Duration**: This metric reflects the mean duration of individual gaze fixations recorded while a stimulus was displayed, computed as the average interval between successive fixation onsets (derived from consecutive FPOGS timestamps) within each trial. In the context of our study, it serves as an indicator of the average momentary visual processing allocated per fixation while evaluating a news claim, rather than the cumulative time spent on the stimulus as a whole. Visit Duration is expressed in milliseconds.

- **LPD** (Left Pupil Diameter): this metric captures the diameter of the left pupil throughout in pixels. Pupil dilation is associated with cognitive load, emotional arousal, and mental effort, providing insight into the cognitive and affective processes underlying credibility judgments.

- **RPD** (Right Pupil Diameter): Similarly, this metric records the diameter of the right pupil in pixels. By analyzing both left and right pupil diameter measurements, we can obtain a more robust indicator of pupillary response while accounting for potential measurement artifacts or individual physiological variations.

6.4.10. STATISTICAL ANALYSIS

The sample meets the requirements of the central limit theorem, which ensures that the sampling distribution of the mean approaches normality regardless of the underlying population distribution when $n \geq 30$, thereby enabling the use of parametric statistical tests [140].

Methodological Note on Data Independence: Since the same participants evaluated multiple claims, the independence assumption of the chi-square test may be violated due to within-participant correlation. To address this issue, we fitted a Generalized Estimating Equations (GEE) model with participants treated as clusters and a binomial distribution for the binary outcome (false-negative vs.

correct classification). This approach provides population-averaged estimates while adjusting standard errors for repeated observations from the same participant. The results obtained with the GEE model were consistent with those of the chi-square analysis. To avoid the duplicate, only GEE results are present as both methods obtained the same result.

To account for within-participant dependence, the GEE models were estimated using both an independent and an exchangeable working correlation structure. Results were substantively identical across correlation structures; therefore, results from the exchangeable structure are reported. The software SPSS version 29 has been used for the analysis and Claude.ai for the coding.

6.5. RESULTS

The results below concern only eye-tracking metrics data and are based to the research questions:

RQ3.1: Do topics and the presence of PPE in fake news significantly influence readers' credulity?

To quantify credulity bias, we computed each participant's false-negative responses based on a binary classification (false/real) and then averaged these values across participants. In total, 660 responses were analyzed (30 participants × 30 claims = 900 responses, minus 240 not_sure responses). The 'not_sure' responses were excluded from the confusion matrix to maintain a strict binary classification framework (real vs fake). Since the objective was to quantify the credulity bias specifically through false-negative rates, we focused on definitive misclassifications. 'Not_sure' responses represent a state of cognitive uncertainty rather than a judgment error; including them would dilute the measures of Type I and Type II errors, which are the standard metrics for evaluating detection performance in a confusion matrix. Figure 6.4 presents the confusion matrix, in which false-negative responses amount to 149 out of 344.

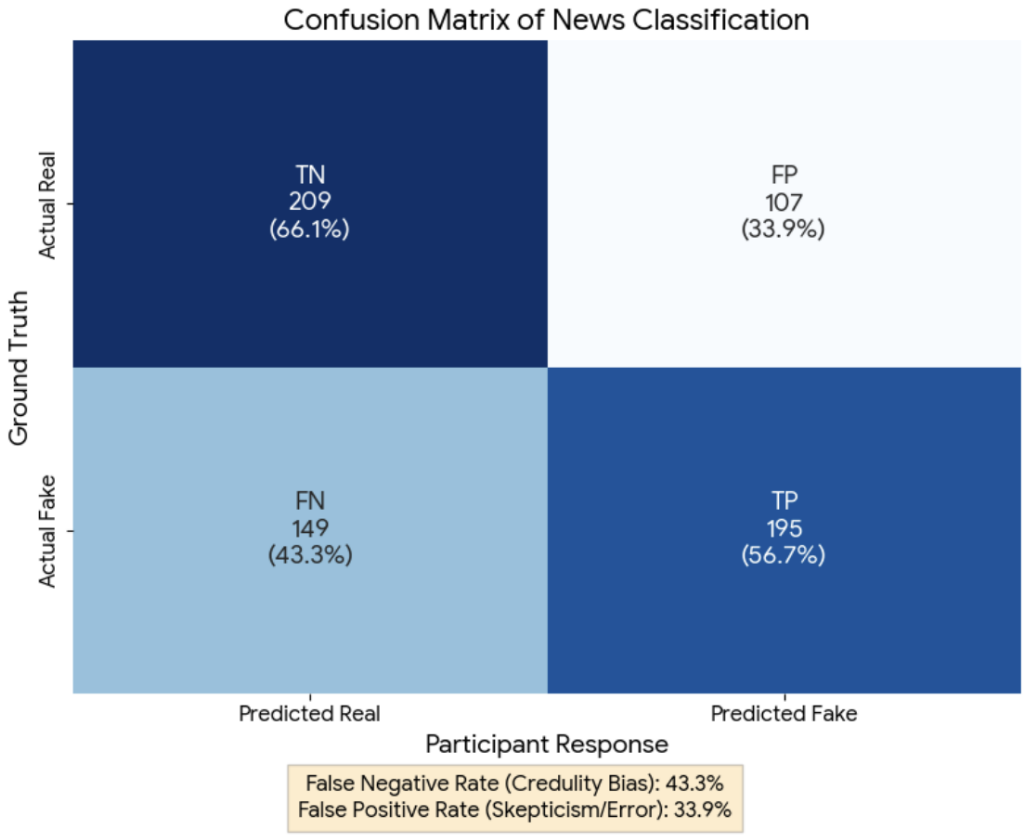


Figure 6.4 : Confusion matrix

We compiled 149 false-negative responses across topics. The results are summarized in Table 6.2. Political claims appeared to be the most deceptive, with 60.19% of classified responses resulting in false negatives (65 out of 108). Environmental claims yielded the largest absolute number of false negatives (N=70), representing 44.03% of classified responses. Finally, science claims showed a lower false-negative rate of 18.18% (14 out of 77).

Table 6.2 : Distribution of False-negative responses (FN) by Topic

Topics	Total Quest.	FN	Total Classified	FN Rate(%)
Environment	7	70	159	44.03
Politics	5	65	108	60.19
Science	3	14	77	18.18
Total	**15**	**149**	**344**	**43.31**

Testing the Role of News Topics (H_{1a}): To test H_{1a}, we used a logistic Generalized Estimating Equations (GEE) model which treats participants as clusters to account for within-subject correlation. Results were substantively identical across correlation structures; therefore, results from the exchangeable structure are reported. Generalized Estimating Equations (GEE) model opted for environmental claims as the reference category.

There was a significant overall effect of Topic on credulity (Wald $\chi^2(2) = 30.198, p < .001$) indicating that susceptibility to deception is not uniform across topics. Relative to Environment (reference category), Politics claims had 1.9 times higher odds of being misclassified as true. (OR=1.902, 95% CI [1.247, 2.899], Wald's $\chi^2 = 8.921, p = .003$), whereas the odds of misclassification were reduced by 72% for Science claims compared to Environment claims (OR=0.280, 95% CI [0.147, 0.533], Wald's $\chi^2 = 15.022, p < .001$). The results confirm H_{1a}. Summary statistics are reported in Table 6.3.

Table 6.3 : Logistic GEE Results for Topic (reference = Environment)

Contrast	OR	95% CI	Wald's χ^2 (1)	p
Politics vs Environment	1.902	[1.247-2.899]	8.921	.003
Science vs Environment	0.280	[0.147-0.533]	15.022	$<$.001

The substantial variation in false-negative rates across individual questions (range: 12.50% to 95.24%) indicates that content-specific features beyond topic category significantly influence believability. Questions 11 (politics, 95.24%) and 15 (environment, 85.19%) proved exceptionally deceptive, while question 3 (environment, 12.50%) was readily detected despite being in the same topic category. This within-topic variability motivates our subsequent investigation into specific proper nouns of political figures (PPE) such as Obama, and Biden.

Testing the Influence of PPE (H_{1b}): To test (H_{1b}), we estimated a logistic GEE whether to examine the presence of PPE increased participants' tendency to judge fake news as true. This primary analysis includes all responses to fake-news claims. The dependent variable was whether a fake-news

claim was judged as true, with responses coded as credulous when participants answered "real" to a fake claim and skeptical when they correctly rejected it.

We first extracted PPE within each claim. Of the 15 fake claims presented to participants, 10 (66.7%) contained no PPE, while 5 (33.3%) explicitly mentioned recognizable PPE. This resulted in 344 total participant responses to fake claims: 236 to claims without PPE and 108 to claims with PPE. Thus, among those cases where participants answered "real" when the claim was actually fake (false negatives), 59 false-negative responses contained PPE. The credulity bias of 43.3% (149/344) represents the overall proportion of participants who believed the fake news as shown in Table 6.4.

Table 6.4 : Participant Responses to Fake News by PPE Presence

	Believed (Credulous)	**Not Believed (Skeptical)**	**Total Responses**
PPE	59	49	108
No PPE	90	146	236
Total	**149**	**195**	**344**

As illustrated in Figure 6.5, the presence of PPE increased participants' tendency to believe fake news. When fake news contained PPE, participants exhibited a credulity bias rate of 54.63% (59/108), compared to 38.14% (90/236) for claims without PPE. This represents an absolute increase of 16.5% in susceptibility to deception.

The results indicated that PPE presence significantly increased the probability of misclassifying a fake news item as true (OR $= 2.693$, 95% CI [1.755, 4.133], Wald's $\chi^2(1) = 20.543$, $p < .001$). The results confirm H_{1b}.

RQ3.2: Do PPE influence visual attention and cognitive engagement during the processing of fake news?

Methodology: To investigate the potential physiological mechanisms underlying the increased credulity toward fake news containing PPE, we conducted an additional GEE analysis including all

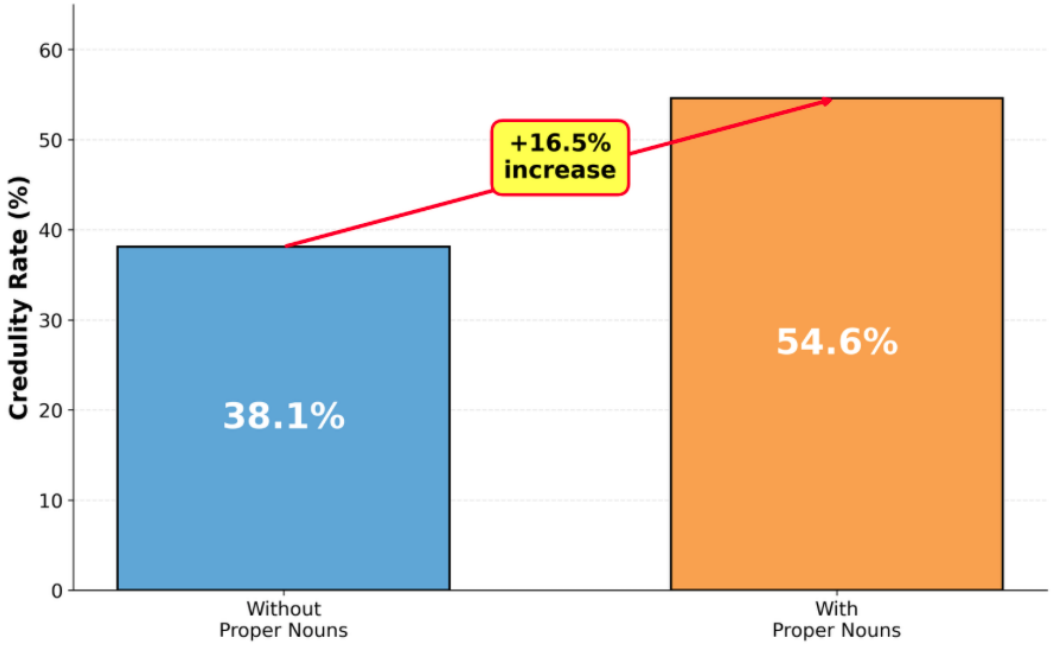


Figure 6.5 : Impact of PPE on Fake news credulity

fake-news trials with valid eye-tracking data (N=330) to determine whether PPE also influenced visual processing beyond response accuracy.

Analysis of Visual Attention and Visit Duration (H_{2a}): To test H_{2a}, we estimated linear GEE model for each continuous metric, using participant as the clustering variable and an exchangeable working correlation structure. Each eye-tracking metric (FPOG, LPD, RPD) was modeled as a function of PPE presence (with vs. without PPE). The behavioral dataset included 344 responses to fake-news claims. For the eye-tracking analysis, 330 valid trials were retained after excluding trials with missing gaze data. Table 6.5 summarizes the results.

This analysis showed that individual gaze fixations were significantly shorter in duration when participants viewed fake claims containing a PPE than when viewing fake claims without a PPE (95% CI [-88.02, -59.48], Wald's χ^2 (1) = 102.63, $p < .001$). This pattern reflects faster momentary visual processing per fixation in the presence of a PPE mention, rather than a difference in the overall time spent engaging with the claim as a whole.

This physiological evidence is consistent with the notion that PPE act as a rapidly-processed visual cue, prompting faster perceptual encoding at the point of fixation, a pattern consistent with more heuristic-based processing. Thus, H_{2a}, is supported.

Table 6.5 : Linear GEE results for Eye-Tracking Across All False-News Trials ($N = 330$)

Measure	With PPE (SD)	w/o PPE (SD)	95% CI	Wald $\chi^2(1)$	p
LPD (camera px)	12.04 (2.02)	11.87 (2.02)	[-0.12, 0.46]	1.350	.246
RPD (camera px)	12.56 (2.20)	12.44 (2.14)	[-0.22, 0.46]	0.450	.502
FPOG(X) (normalized 0–1)	0.36 (0.09)	0.37 (0.08)	[-0.03, 0.01]	1.510	.220
FPOG(Y) (normalized 0–1)	0.21 (0.06)	0.21 (0.07)	[-0.01, 0.01]	0.000	1.000
Visit Duration (ms)	192.32 (147.95)	266.07 (168.53)	[-88.02, -59.48]	**102.630**	**<.001**

Pupillary Response and Cognitive Load (H_{2b}): Regarding H_{2b}, we examined the Left and Right Pupil Diameters (LPD and RPD) as proxies for cognitive effort. As illustrated in Table 6.7, our analysis found that there is no significant difference in mean pupil diameter between false negatives containing PPE and those without, rejecting H_{2b}.

6.6. DISCUSSION

Role of the topic. As presented in Table 6.2, 60% of the false negatives concern political fake news. This high susceptibility rate suggests that political content exploits existing biases, partisan alignments, or emotional engagement that interfere with cognitive processing. According to the CISION Global State of the Media report, the areas where fake news circulated the most between 2020 and 2025 are mainly related to health and politics[144]. Notably, question 11 (Obama) achieved a 95.24% (Figure 6.6) false-negative rate or only one participant classified correctly fake, indicating near-universal acceptance of this particular political claim. Indeed, Pennycook and Rand [145] argue that susceptibility to political disinformation is better explained by inattention to accuracy and limited analytic reasoning. According to the authors, people do not actively seek to believe disinformation; instead, they evaluate credibility based on superficial cues, especially when they are not paying attention to accuracy. People use heuristics, which are cognitive shortcuts that allow them to quickly judge

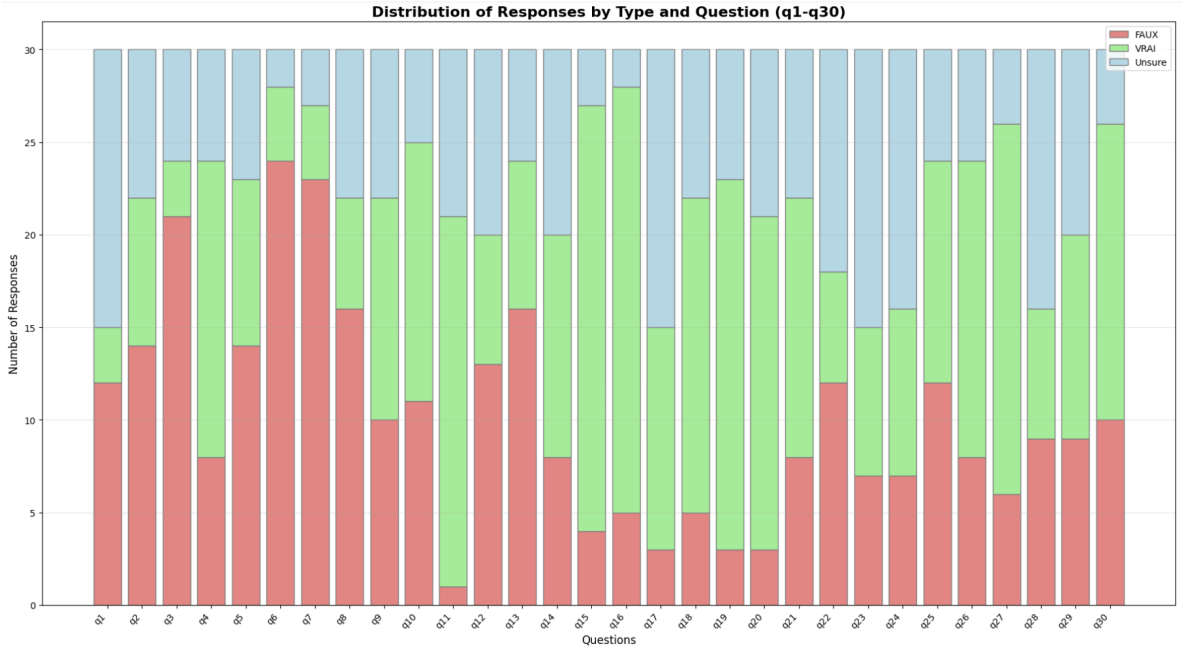


Figure 6.6 : Participants' responses (real, fake or not_sure) by claims

whether a piece of information seems true or plausible, without engaging in a thorough analysis of its factual content.

The second place of credulity bias concerns environmental fake news (44% - see Table 6.2). Particularly, question 15 showed an 85.19% false-negative rate (Figure 6.6), the second-highest individual susceptibility. The prevalence of environmental fake news may reflect the complexity of environmental issues. Indeed, environment and fake news are deeply intertwined, with deliberate campaigns by vested interests (like fossil fuels) and accidental sharing by users creating confusion, delaying climate action, undermining science, etc. [146]. Finally, the science lower susceptibility suggests that participants may apply knowledge to scientific claims.

Role of the PPE. The substantial variation in false-negative rates across individual questions indicates that content-specific features beyond topic category influence believability. As politic topic is the most cognitive processing influence, we decided to dive deeper with the PPE. Indeed, the

presence of PPE significantly increased participants' susceptibility to deception. Fake news creators who invoke recognizable authority figures can effectively exploit cognitive shortcuts to enhance the perceived legitimacy of fabricated claims. This finding aligns with dual-process theories of cognition, suggesting that familiar names trigger heuristic processing rather than analytical evaluation. When encountering statements attributed to or concerning known authorities, participants may engage in minimal verification, assuming that information connected to legitimate entities is more likely to be accurate [71].

Additional analysis on real-news claims. We examined whether PPE effects extend to real-news trials by analyzing false-positive responses (cases where true claims were incorrectly judged as false). This analysis was not intended to serve as a statistical control for claim plausibility or emotional wording, but rather to assess whether the observed PPE-related effects generalize beyond deceptive content to truthful information.

When real claims contained a PPE, participants not correctly identified them as true (Not Believed) at a rate of 32.69% (17/52), compared to 34.09% (90/264) for real claims without PPE, an absolute difference of -1.40 percentage points as shown in Table 6.6.

Table 6.6 : Participant Responses to Real News by PPE Presence

	Believed (Credulous)	**Not Believed (Skeptical)**	**Total Responses**
PPE	35	17	52
No PPE	174	90	264
Total	**209**	**107**	**316**

To account for within-subject dependence, we estimated a logistic GEE model, analogous to the model used for fake claims. The results indicated that the presence of a PPE did not significantly affect correct classification of real news claims (OR $= 1.065$, 95% CI $[0.566, 2.005]$, Wald's $\chi^2(1) = 0.038$, $p = .846$). These findings complement and extend the primary GEE results by showing that the PPE effect is not a general credibility amplifier but a selective bias that operates specifically in the context of deception.

Role of the visit duration. The combination of increased credulity rates and substantially shorter individual fixation durations for fake news containing PPE indicates a qualitative shift in cognitive processing strategies rather than a simple reduction in attention. The 27.8% decrease in mean fixation duration for all fake news, and 42.5% decrease for false-negative responses, suggests that the visual and cognitive processing engaged at each fixation became more heuristic, faster and less deliberative, when participants encountered PPE.

From an attentional perspective, shorter individual fixation durations are indicative of reduced cognitive resource allocation at each point of visual inspection. Rather than dwelling analytically on each element of the claim, participants appear to have processed the content more rapidly at the level of individual fixations. This pattern is consistent with the notion that PPE function as a cognitive shortcut, signaling presumed relevance or credibility and thereby promoting a more heuristic mode of visual processing.

This effect aligns closely with theories of heuristic processing. According to the heuristic–systematic model [143], individuals under low motivation or cognitive effort conditions rely on surface cues rather than message content. PPE, by invoking familiar political figures, likely activate pre-existing schemas that substitute for analytical evaluation. Within the framework of Dual Process Theory [71], PPE introduce a state of cognitive ease, promoting System 1 processing at the expense of System 2 engagement. The substantial reduction in individual fixation duration suggests that this shift toward heuristic processing occurs rapidly, at the level of each visual fixation.

Moreover, the presence of PPE appears to accelerate confirmation bias mechanisms [147]. When PPE align with participants' prior attitudes, the information fits existing mental models and is therefore processed more fluently. This fluency reduces the perceived need for verification, resulting in superficial visual inspection. In this sense, PPE do not merely bias judgment; they short-circuit the evaluation process itself.

Importantly, these findings are consistent with Pennycook and Rand's argument that belief in fake news is driven primarily by insufficient cognitive engagement rather than motivated reasoning [145]. The eye-tracking data provide physiological support for this claim: participants did not appear to evaluate the content and then accept it, but rather failed to engage in extended evaluation altogether. PPE thus operate as a "fast-track" signal that prematurely terminates analytical processing, offering a mechanistic explanation for their strong effect on false-negative responses.

While the lack of significant difference in pupillary response might initially seem counterintuitive (H_{2b}), it provides a critical piece of the cognitive puzzle when viewed alongside the drastic reduction in individual fixation duration (H_{2a}). Pupil dilation is a well-established proxy for mental effort, task difficulty, and the engagement of analytical resources [82]. The fact that pupil diameter remained stable suggests that the presence of PPE did not trigger a "red flag" or increase the perceived difficulty of the task.

These findings highlight a critical vulnerability in human-centered fake news detection. If linguistic elements like PPE can shift readers toward a faster, more heuristic style of visual processing, automated systems cannot rely solely on humans to catch sophisticated fake news. Instead, HAI systems should be designed to detect these linguistic triggers. Moreover, AI-tailored disinformation will exploit human cognitive vulnerabilities with unprecedented precision [148].

Secondary analysis restricted to false-negative responses. To further explore the mechanism underlying this effect, we conducted a secondary analysis restricted to false-negative responses (N=147).

Table 6.7 : Linear GEE result for Eye-Tracking within False Negatives (N=147)

Measure	**With PPE(SD)**	**w/o PPE(SD)**	**95% CI**	**Wald's $\chi^2(1)$**	p
LPD (camera px)	12.12 (1.73)	11.74 (1.93)	[-0.12, 0.24]	.001	.981
RPD (camera px)	12.64 (1.96)	12.37 (1.97)	[-0.27, 0.17]	0.465	.495
FPOG(X) (normalized 0–1)	0.36 (0.07)	0.39 (0.06)	[-0.04, -0.01]	**8.285**	.004
FPOG(Y) (normalized 0–1)	0.21 (0.05)	0.20 (0.06)	[-0.02, 0.01]	0.03	.86
Visit Duration (ms)	159.66 (126.94)	277.74 (147.04)	[-139.03, -83.72]	**62.30**	<.001

The results of this analysis, shown in Table 6.7, revealed two significant differences. A linear GEE model indicated that:

-FPOG(X): the presence of PPE significantly elicited a horizontal gaze position further from the screen center (95% CI $[-0.04, -0.01]$, Wald's $\chi^2(1) = 8.285$, $p = .004$).

-Visit Duration: the presence of PPE significantly shortened individual fixation duration (95% CI $[-139.03, -83.72]$, Wald's $\chi^2(1) = 62.30$, $p < .001$).

This complementary analysis showed an amplified reduction of 42.5% in mean fixation duration when participants accepted fake news, suggesting that PPE may facilitate a faster, more heuristic processing mode during deceptive judgments. This reduction observed in false negatives ($p < .001$) was absent in false positives ($p = .115$), suggesting that the cognitive shortcut triggered by PPE is specific to the processing of fake news. When participants failed to identify true claims, PPE did not shorten fixation durations nor alter pupillary response.

Secondary analysis restricted to false-positive responses. To extend the primary GEE results by showing that the PPE effect is not a general credibility amplifier but a selective bias that operates specifically in the context of deception, we conducted an linear GEE model analysis focusing on true news claims that participants incorrectly classified as false, examining the same eye-tracking metrics (FPOG, LPD, RPD) with respect to PPE presence. The results are summarized in Table 6.8.

The results revealed a strikingly different pattern from that observed in false negatives. None of the fixation duration or pupillary metrics were significantly affected by PPE presence. The only significant effect emerged for the vertical gaze position FPOG(Y) (95% CI $[-0.052, -0.007]$, Wald's $\chi^2(1) = 6.400$, $p = .011$), reflecting a subtle downward shift in fixation when PPE were present, with no counterpart in the false negative condition.

Table 6.8 : Linear GEE result for Eye-Tracking Metrics within False Positives (N=106)

Measure	With PPE(SD)	w/o PPE(SD)	Wald's $\chi^2(1)$	p
LPD (camera px)	11.64 (2.33)	11.51 (1.71)	0.005	.95
RPD (camera px)	12.17 (2.77)	12.01 (1.90)	0.008	.93
FPOG(X) (normalized 0–1)	0.34 (0.13)	0.38 (0.1)	0.80	.37
FPOG(Y) (normalized 0–1)	0.21 (0.08)	0.23 (0.1)	6.4	**0.01**
Visit duration (ms)	414.28 (219.53)	345.31 (206.35)	2.48	.12

Role of the heatmaps. The heatmaps provide qualitative visual support for the possibility that PPE may serve as salient cues. To visualize the distribution of visual attention across the textual claims, we generated the heatmaps from the raw `FPOGX/FPOGY` fixation coordinates exported by the Gazepoint system.

The color intensity represents the spatial density of fixation points aggregated across all 30 participants for each claim. Warmer colors (red) indicate areas of high visual interest, while cooler colors (yellow) indicate areas with less visual attention. Heatmaps with PPE presence within fake news are included. In Figure 6.7, the attention is primarily concentrated around the name *Donald Trump*, while the name *Hitler* attracts virtually no interest. In Figure 6.9, the attention is even more strongly focused on the name *Trump*. Finally, in Figure 6.10, the attention is directed toward the name *Biden*, but less markedly so.

To formally evaluate these visual patterns, we addressed the AOI-based measures around PPE terms (e.g., fixation count, dwell time).

The methodology comprised the following steps.

-Identification of PPE-containing in fake news claims (Figures 6.7, 6.8, 6.9, 6.10).

-AOI definition. To precisely define the area of interest (AOI) corresponding to each PPE term, we extracted the exact video frame from the Gazepoint screen recording (1920 × 1080 resolution) at the moment each claim was displayed. Optical character recognition (OCR) was then applied to obtain

the exact pixel coordinates of each PPE term, which were subsequently converted to the normalized coordinate system (0–1) used by Gazepoint's `FPOGX/FPOGY` gaze coordinates.

-Coordinate verification. We verified that no axis transformation was required by comparing the spatial locations of observed fixations with the corresponding text positions in the extracted video frames, confirming accurate coordinate alignment between gaze data and AOI locations.

-Calibration check and AOI adjustment. As a calibration check, we compared the centers of the AOI bounding boxes with the actual locations of nearby fixations. A small spatial offset (approximately 2–4% of screen width) was observed during manual verification, potentially reflecting calibration drift or natural fixation dispersion.

-Normalization of eye-tracking measures. Because PPE terms occupied only one or two words, whereas the remainder of the claim text contained between 26 and 48 words, fixation counts and dwell times were normalized by the number of words in each region (i.e., per-word density measures). This normalization enabled a more equitable comparison between the PPE region and the remainder of the text. Word-count normalization was selected because the objective was to compare attention allocated to linguistic units rather than to visual area, given the substantially smaller spatial extent of PPE terms relative to the remainder of the claim text.

-Computation of AOI measures. A bounding box encompassing the entire claim text was first defined for each claim. For each participant, fixation count and dwell time were computed separately within (a) the full claim text-box AOI and (b) each PPE AOI. Fixation count and dwell time attributable to the remainder of the text were then obtained by subtracting the PPE AOI values from the corresponding text-box totals. Per-word density measures were derived by dividing fixation count and dwell time by the number of words in the PPE region and in the remainder of the text, respectively.

We estimated linear GEE models, with participant as the clustering variable and an exchangeable working correlation structure, modeling fixation density and dwell-time density as a function of region (PPE vs. remainder of text). The analytic dataset contained 240 observations (30 participants × 4

PPE-containing claims × 2 regions). We conducted supplementary analyses in which AOI boundaries were symmetrically expanded by 0.01, 0.02, 0.03, and 0.05 normalized units beyond the exact word boundaries. The GEE models were re-estimated at each margin to assess whether conclusions were robust to reasonable variation in AOI precision. It should be noted that only the 0.01–0.02 margins fall within the device's nominal gaze-accuracy range. The 0.03 and 0.05 margins were included to characterize the broader trend.

The results show that both effect size and statistical evidence increased monotonically with margin width, suggesting a possible attentional preference for the PPE term that becomes detectable only once natural gaze dispersion is accounted for. Although this trend did not reach conventional significance at any tested margin, the dwell-time density effect approached significance at the most permissive margin ($p = .064$ at 0.05) in Table 6.9, suggesting that a modest effect may be present.

Table 6.9 : Sensitivity Analysis Across AOI Margins

	Fixation Density		**Dwell Density**	
AOI Margin	*b*	*p*	*b*	*p*
0.01	0.009	.974	0.017	.857
0.02	1.020	.282	0.265	.276
0.03	2.200	.181	0.593	.167
0.05	4.862	.087	1.330	.064

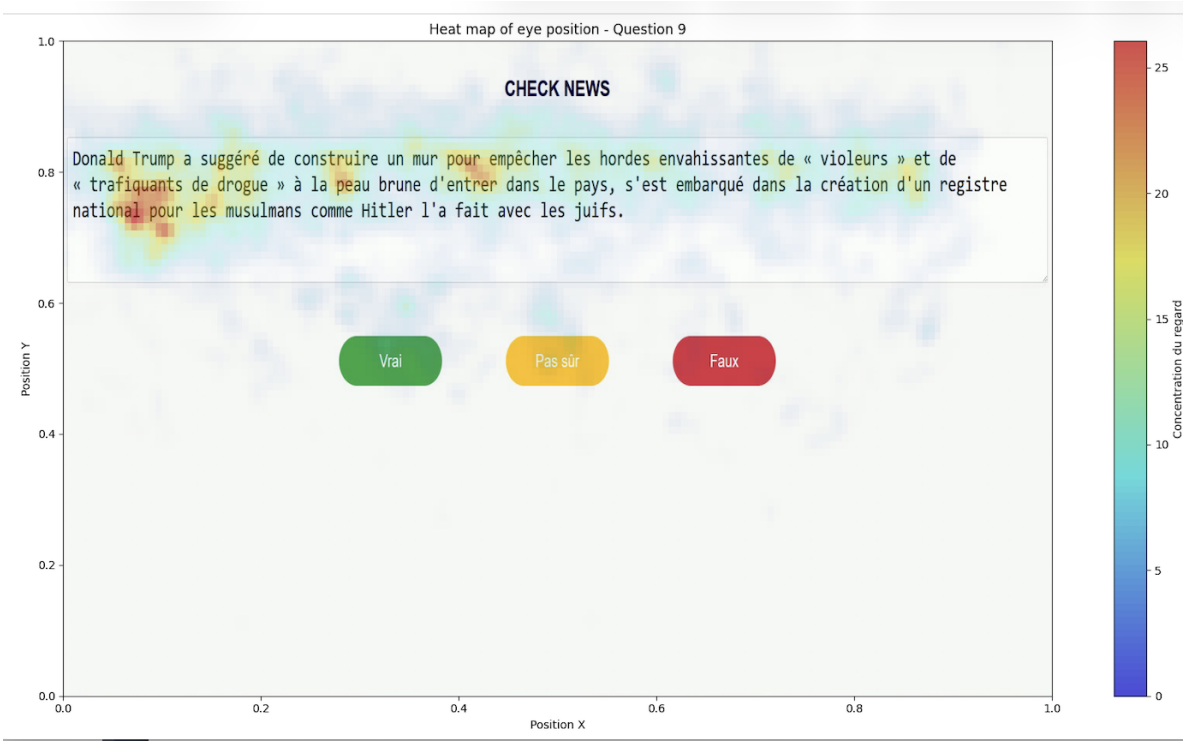


Figure 6.7 : Eye-Tracking Heatmap with PPE "Donald Trump"

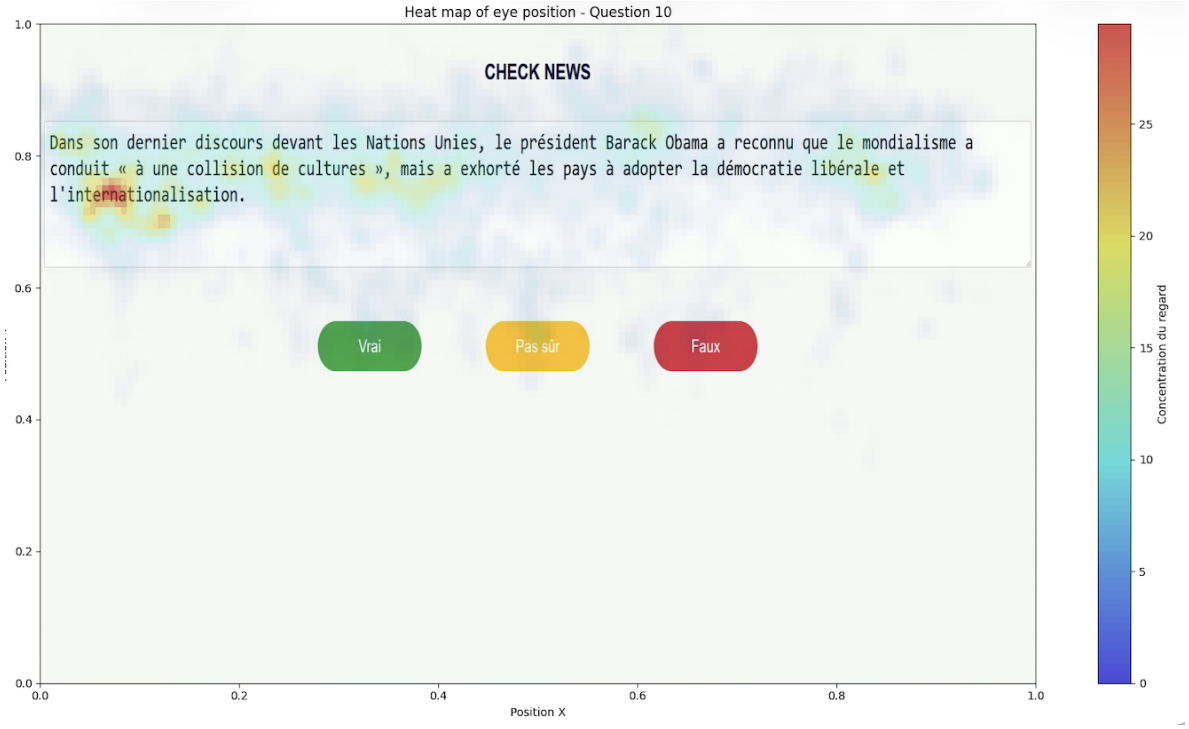


Figure 6.8 : Eye-Tracking Heatmap with PPE "Obama"

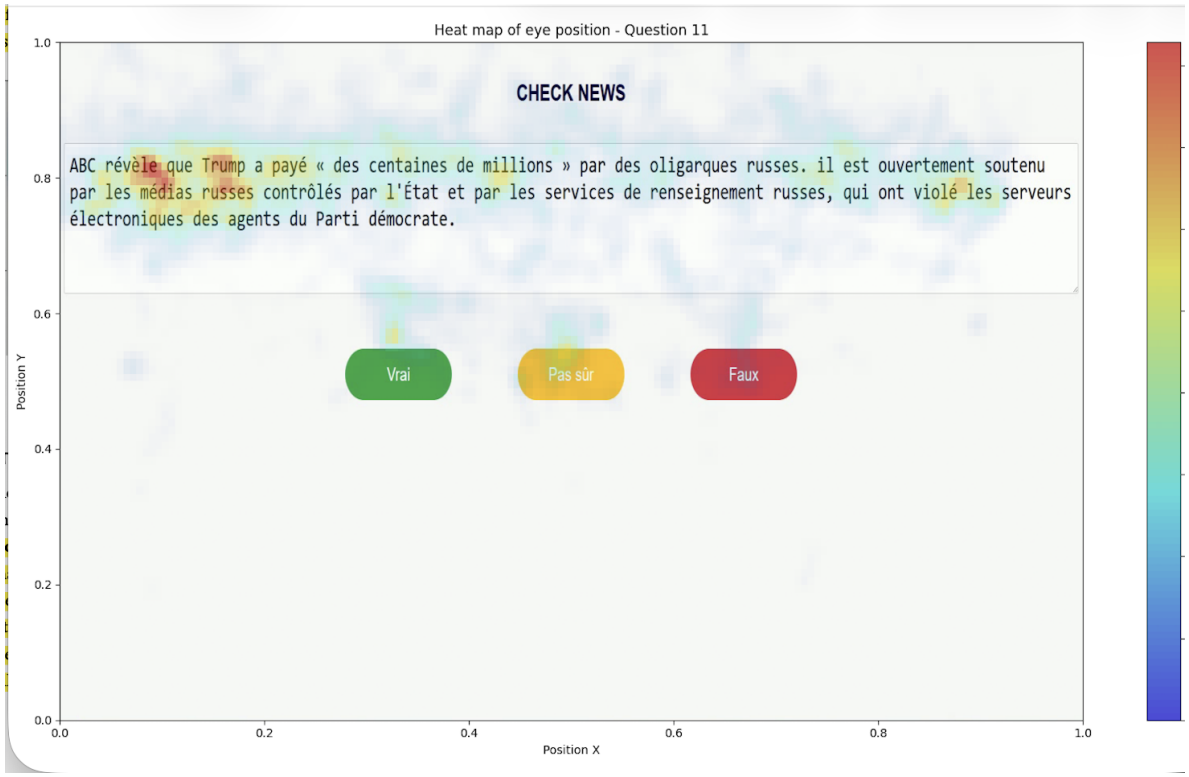


Figure 6.9 : Eye-Tracking Heatmap with PPE "Trump"

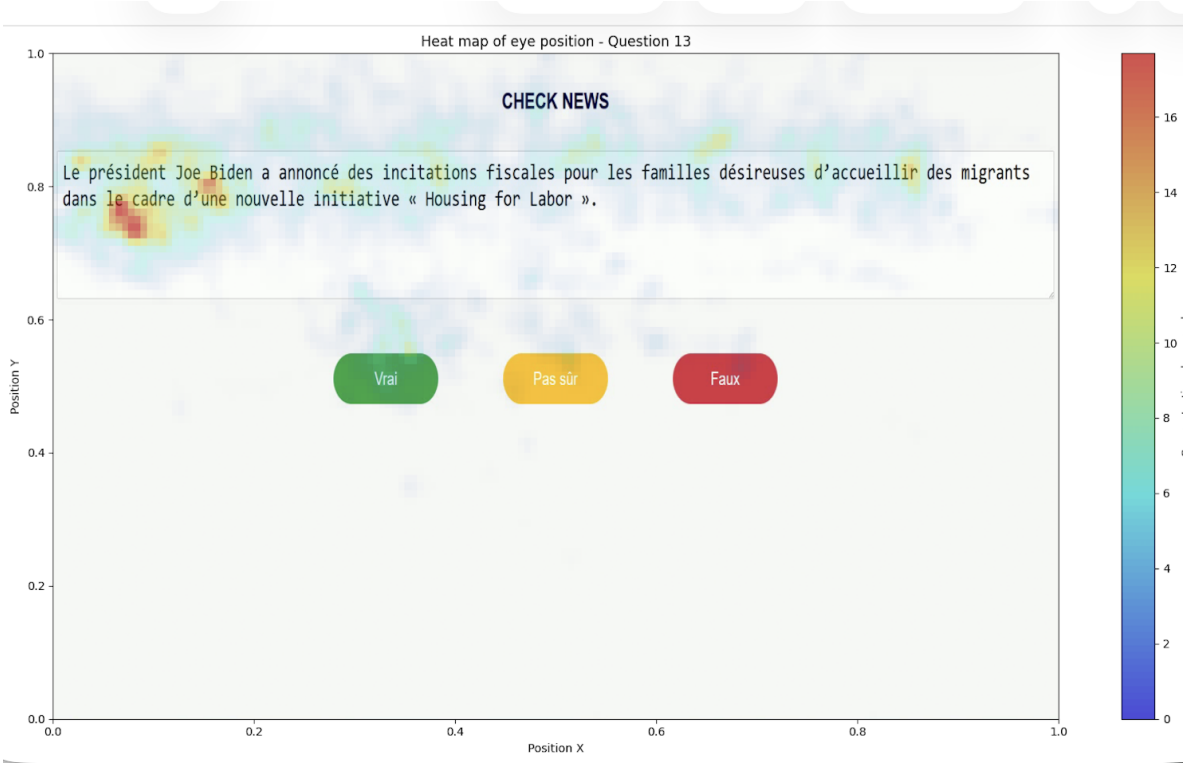


Figure 6.10 : Eye-Tracking Heatmap with PPE "Biden"

6.7. THREATS TO VALIDITY

While this study enabled discovery of credibility patterns across varied content, it introduces specific validity considerations that we acknowledge and address.

Construct Validity concerns whether our measurements accurately reflect the concepts we intend to study. The primary threat was whether our eye-tracking metrics truly captured the effect of PPE, or merely reflected other phenomena such as familiarity effects or attentional artifacts. We addressed this through methodological triangulation: eye-tracking metrics converged on the same pattern showing reduced systematic processing when PPE were present, as well as eye-tracking heatmap analysis. We also added an analysis of real-news claims showing that PPE presence did not significantly

affect the classification of truthful claims. This additional analysis helps clarify that the observed PPE-related credibility effect is not simply a general response tendency across all claims.

Internal Validity concerns the extent to which our study can establish a causal relationship between the variables under investigation. The selection of American political figures for our PPE stimuli was a deliberate design choice rather than an incidental feature of the stimulus set: research on source credibility and the credulity bias suggests that individuals hold pre-existing, differentiated trust priors toward specific public figures (e.g., a figure widely associated in public discourse with misinformation, such as Donald Trump, versus a figure more commonly perceived as credible, such as Barack Obama), and PPE was designed to leverage this naturally occurring credibility gradient rather than to introduce a neutral, content-free tag. This design choice, however, means that the observed PPE effect cannot be fully separated from participants' pre-existing attitudes toward the specific named figures used: what we interpret as a "PPE effect" partly reflects the credibility (or lack thereof) already attached to these particular figures, rather than a generic effect of citing any public figure. Because participants' political orientation and prior attitudes toward these figures were not measured, we cannot determine the extent to which individual differences in these priors moderated the observed effect, nor generalize it beyond this specific set of highly salient, polarizing figures.

Future work should directly measure participants' political orientation, prior attitudes toward the specific public figures used, and familiarity with them as covariates, and should ideally cross figure identity with a neutral-source condition, in order to disentangle a generic PPE/source-citation effect from figure-specific credibility priors.

Reliability Validity concerns the possibility of replicating our study's findings. Our study was conducted in the context of a neutral laboratory that differs from how participants are used to reading news claims in their everyday life. Often, readers have the opportunity to cross-reference information to verify claims. However, this "controlled" approach was a deliberate design choice. This study aims to isolate and capture the initial cognitive response, the "first sign" of discernment, before extended research behaviors intervene. Participants in a lab are aware they are being monitored, which can

trigger a Hawthorne effect, where they might exert more (or different) effort than they would during a casual scroll through a social media feed [149].

External Validity refers to the extent to which the findings of a study can be generalized beyond the experimental setting. Several factors may limit the generalizability of our results, including the restricted number of claims in certain topical categories (e.g., health), the limited set of PPE, the demographic composition of participants, and the predominance of American political figures in the stimulus material. Because only a limited number of fake-news claims contained PPE, the observed association may partly reflect item-specific characteristics of those claims, such as wording, topic, emotional tone, or perceived plausibility. Consequently, the findings should be interpreted as reflecting responses to highly salient public figures within this specific cultural context and should not be assumed to generalize to all populations, named entities, or political environments.

To address the unequal distribution of claims across topics (7–5–3), we conducted bootstrap sensitivity analyses using balanced resampling designs across 10,000 iterations. The results remained stable and statistically consistent with the primary analyses, suggesting that the observed effects were not driven by topic imbalance alone.

Recruitment was constrained by the laboratory-based eye-tracking protocol, which required in-person participation with fixed equipment and resulted in a relatively small sample drawn from a single university. This has an important consequence beyond sample size: our participants formed a highly educated population (university students and/or staff), whose reading habits, media literacy, and critical-thinking skills likely exceed those of the general population. Educational attainment is a well-documented predictor of resistance to disinformation, and it plausibly shaped how our participants comprehended and evaluated the fake-news claims. Consequently, the effects we observed may be attenuated, or take a different form, in less-educated or more heterogeneous populations, and should not be assumed to generalize beyond an educated, laboratory-based sample.

6.8. CONCLUSION AND FUTURE WORK

In conclusion, this chapter examined a novel perspective by examining the role of news topics and the presence of PPE in shaping both behavioral susceptibility to fake news and visual processing patterns. Our empirical study, involving 30 participants who each evaluated 30 claims under continuous eye-tracking, revealed important insights.

The findings provide strong evidence that both the **topic** of a news item and the **presence of Political Person Entities (PPE)** significantly influence readers' credulity. The results show that susceptibility to fake news varies across topics and that PPE increase the likelihood of misclassifications. These effects are consistent with the cognitive heuristic framework, according to which familiar or salient cues reduce analytical scrutiny and promote intuitive acceptance.

The eye-tracking analyses reveal that fake news containing PPE elicit significantly shorter fixation durations, indicating more heuristic-based processing. However, no significant effect was observed on pupillary responses, suggesting that PPE influence attention allocation more strongly than cognitive load. Moreover, the heatmaps provide qualitative visual support suggesting that a modest effect may be present ($p = .064$ at 0.05).

Overall, the results support the sub-hypothesis 3 that Political Person Entities (PPE) and topic used in fake news can contribute to misclassifying it as real news.

Future work should aim to strengthen causal inference by systematically vary identical claims contain PPE or generic references (e.g., "A president" vs. "President Trump"). Moreover, although the current sample size was sufficient to detect robust effects, larger and more diverse participant samples would enable more fine-grained analyses of individual differences and distinctions between different categories of proper nouns (e.g., political figures vs. institutions).

To facilitate replication and future research, the dataset is available at https://anonymous.4open.science/status/Eye-tracking-Fake-News—Study-2876.

CONCLUSION

This dissertation addressed the growing challenge of fake news in contemporary digital environments by investigating how Explainable Artificial Intelligence (XAI) can support more effective and cognitively aligned mitigation strategies. Motivated by the increasing industrialization of information manipulation, the widespread deployment of generative AI, and persistent human cognitive vulnerabilities, this research examined fake news detection through a multidisciplinary lens combining deep learning and XAI, LLMs, and cognitive science.

Ultimately, the findings presented in this dissertation is to validate our **General Hypothesis, which consists of AI-based and cognitive approaches including explainable artificial intelligence (XAI), large language models (LLMs), and cognitive eye-tracking analysis improve fake news mitigation**. The validation of this overarching hypothesis was achieved through the systematic evaluation of three distinct sub-hypotheses, corresponding to the major contributions of this dissertation.

Detecting and explaining fake news remains a fundamentally challenging problem. This difficulty is further amplified by the emergence of advanced generative AI systems, which enable the production of increasingly realistic and persuasive synthetic content, making fake news harder to identify even for human observers. In this context, individuals are often left without reliable external verification tools and must rely primarily on their own cognitive judgment and limited contextual cues to assess the credibility of information.

The situation is further exacerbated by the structure of social media ecosystems, where information is rapidly disseminated, sources are heterogeneous, and the distinction between reliable and unreliable content is often blurred. As a result, fake news can spread widely before verification occurs,

and users frequently lack the means to confidently assess source credibility or factual correctness in real time.

7.1. VALIDATION OF SUB-HYPOTHESIS 1

The first contribution addressed Sub-Hypothesis 1 that consists of augmenting a BERT-based architecture with contextual metadata (e.g., number of followers, hashtags, engagement metrics), combined with SHAP-based feature attribution, improves both predictive fake news performance and explainability.

BERT was fine-tuned on an 80/20 split of 5,000 COVID-19 records, including text and metadata. The remaining 1,000 records were used to calculate performance metrics. The results show that detection performance can be significantly improved by integrating textual and contextual information within a unified Transformer-based architecture, like BERT. Moreover, while SHAP improves transparency by highlighting influential metadata signals. **Overall, these findings support Sub-Hypothesis 1**.

However, these improvements should be interpreted in light of inherent limitations of static Transformer-based architectures in dynamic disinformation environments. Models such as BERT are trained on historical corpora and therefore remain constrained by temporal bias and limited adaptability to emerging narratives and evolving deception strategies. In the context of fake news, where content rapidly evolves and exhibits strong distribution shifts over time, this limitation reduces the model's practical robustness in real-world deployment scenarios.

In addition, while SHAP provides useful post hoc feature attribution, it does not capture causal mechanisms underlying fake news. It highlights feature importance rather than interpreting the narrative structure or factual inconsistencies that render content deceptive. Consequently, although both predictive performance and transparency are improved, the model remains limited in its ability to provide semantically grounded explanations of fake news.

Despite the improvements achieved through the integration of contextual metadata and SHAP-based feature attribution, the findings of Sub-Hypothesis 1 also reveal fundamental limitations of

discriminative architectures in addressing the broader problem of fake news mitigation. In particular, while Transformer-based models such as BERT are effective at learning statistical associations between textual and contextual signals, they remain inherently constrained in their ability to capture the evolving and adversarial nature of fake news.

Furthermore, although post hoc explanation methods such as SHAP improve transparency by identifying influential features, they do not provide semantic or causal accounts of why content is misleading. This creates a gap between prediction and understanding: the model can indicate what drives a classification decision, but not why the content is deceptive in a human-interpretable sense.

In sum, BERT's architecture makes it well suited for both binary classification (true or fake) and downstream clustering or semantic analysis, such as theme classification of messages, and its parameters are comparatively easy to fine-tune through standard backpropagation. However, these strengths remain confined to surface-level pattern recognition: BERT does not identify the narratives, entities, or the roles these entities play within information manipulation campaigns. The dynamic characteristics of fake news, including rapid narrative shifts, emerging events, and strategic content manipulation, introduce distributional changes that static models are not explicitly designed to handle.

These limitations motivate a shift toward more expressive architectures capable not only of classification but also of generating structured explanations grounded in external knowledge and implicit reasoning processes. This transition forms the basis of Sub-Hypothesis 2, which investigates whether large language models, through zero-shot prompting, can better support both fake news detection and explanatory generation.

7.2. VALIDATION OF SUB-HYPOTHESIS 2

The second contribution addressed Sub-Hypothesis 2 that consists of evaluating the zero-shot prompted LLMs to effectively detect fake news and generate explanations that contribute to fake news mitigation. Three open-source LLMs were instructed with a prompt asking them to classify

content as fake or real and justify their answers. The results show that all tested models achieved modest performance on the Fin-Fact dataset, with accuracy scores marginally above chance baseline ($\approx$ 0.54–0.55) for both true and fake news. To evaluate the quality of the generated explanations, LLMs achieve good BERTScore-F1 values. Moreover, LLaMA4 shows effective use of an implicit chain-of-thought reasoning process.

Overall, these findings partially support Sub-Hypothesis 2. While the generated explanations showed semantic alignment with reference explanations, the results suggest that those LLMs remain limited in producing sufficiently transparent and fact-grounded explanations required for specialized fake news mitigation tasks.

These findings point to the need for integrating explainable AI (XAI) techniques into LLM-based fake news detection pipelines, so that generated explanations are explicitly anchored to verifiable facts and traceable to supporting sources, rather than relying solely on the model's implicit reasoning [150].

Beyond their predictive performance, LLMs offer an additional operational advantage that is particularly relevant for fake news analysis in real-world settings. Unlike discriminative architectures such as BERT, which primarily focus on classification tasks, LLMs enable a more flexible and structured form of narrative analysis through instruction-based prompting.

In practical analytical workflows, this capability allows analysts to design prompt-driven scripts that guide the model toward specific dimensions of content interpretation, such as *narrative reasoning*, rhetorical analysis, and *cognitive framing* embedded within social media messages. Consequently, LLMs provide a complementary analytical layer that supports qualitative interpretation and reasoning over content, which cannot be easily captured by static embedding-based models.

7.3. VALIDATION OF SUB-HYPOTHESIS 3

The third contribution addressed Sub-Hypothesis 3 that the cognitive eye-tracking analysis explains how people are influenced by fake news. This exploratory study examined topics and the

presence of PPE (e.g., well-known political figures such as Trump, and Obama) in fake news and showed that they can bias perception and contribute to its misclassification as real news.

Thirty participants were asked to classify 30 claims while eye-tracking recorded their eye movements. The results show that PPEs and specific topics, attract disproportionately high fixation durations. The PPE function as cognitive heuristics, biasing user perception and significantly increasing the likelihood of misjudging fake news as real. **Overall, these findings support the sub-hypothesis 3.**

However, while eye-tracking provides strong evidence of attentional bias, it does not fully explain the underlying narrative mechanisms that trigger such cognitive effects. In this context, LLMs can be used as a complementary analytical tool to bridge the gap between observed cognitive behavior and textual structure.

By leveraging instruction-based prompting, LLMs can be employed to decompose content into narrative components, such as framing strategies, authority cues, emotional valence, and implicit persuasion techniques associated with PPEs. This enables a more explicit mapping between discourse structures and cognitive responses, offering a richer interpretation of why certain entities systematically attract attention and bias judgment in fake news environments.

7.4. SYNTHESIS

Taken together, the outcomes of these three sub-hypotheses, whether fully or only partially supported, converge toward the central thesis of this dissertation: effective fake news mitigation requires combining large language models (LLMs), explainable AI (XAI), and cognitive eye-tracking analysis, rather than relying on any single methodological perspective.

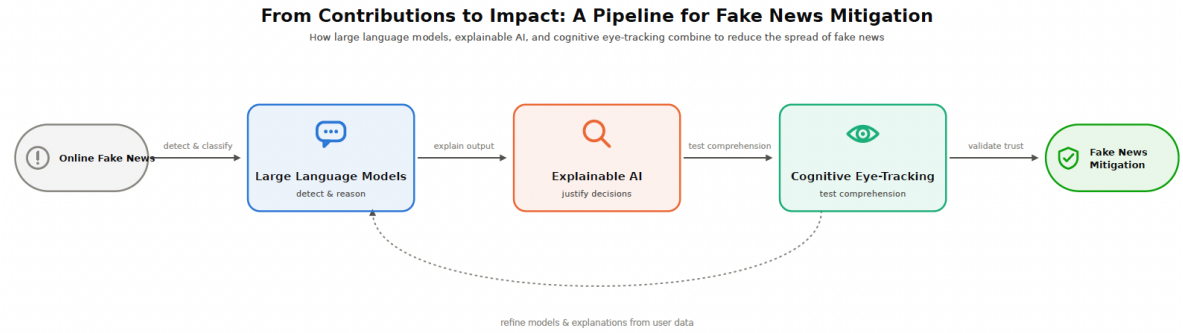


Figure 6.11 : Links between XAI, LLMs and Eye-tracking analysis

Figure 6.11 presents a conceptual pipeline that illustrates how these three research contributions interact to mitigate the spread of online fake news. Fake news is first fed into a large language models module responsible for detecting and reasoning about misleading content. The output of this module is then passed to an explainable AI component, which justifies the model's decisions by generating interpretable explanations. These explanations are subsequently evaluated through cognitive eye-tracking, which tests users' comprehension of them. Finally, the pipeline converges on the overarching goal of fake news mitigation, achieved by validating user trust in the system's outputs.

A feedback loop connects the cognitive eye-tracking stage back to the large language models stage, indicating that insights gathered from user comprehension and gaze behavior are used iteratively to refine both the detection models and their explanations, informed by narrative analysis, and discourse strategies underlying information content.

These findings resonate with emerging perspectives in cognitive security and information integrity research, particularly within the Foreign Information Manipulation and Interference (FIMI) framework. As shown in the work of Carley et al. [151, 152], effective mitigation strategies require the joint exploitation of textual signals, contextual metadata, temporal dynamics, and explainable computational models. Within this expanded view, explainability, narrative analysis, and cognitive grounding are not auxiliary components but fundamental pillars of a unified framework for information influence

analysis, one that this dissertation proposes by articulating LLMs, XAI, and cognitive eye-tracking as three complementary and mutually reinforcing dimensions of fake news mitigation.

7.4.1. THREATS TO VALIDITY AND LIMITATIONS

Despite the promising findings obtained across the three contributions, several limitations must be acknowledged. First, some experiments relied on relatively constrained datasets and domain-specific corpora, which may limit the generalizability of the results across broader fake news ecosystems. Second, the evaluation of LLM-generated explanations remains inherently challenging due to the absence of universally accepted semantic interpretability metrics. Third, the eye-tracking and exploratory EEG experiments were conducted on limited participant samples, potentially affecting statistical power and external validity. Finally, the rapid evolution of generative AI systems and disinformation strategies may impact the long-term stability and reproducibility of the proposed approaches. These limitations nevertheless provide important directions for future research and highlight the necessity of continuously integrating computational, explainable, and cognitive perspectives in fake news mitigation research.

7.5. FUTURE WORK

A promising direction concerns the integration of neurophysiological signals, such as EEG, into the study of disinformation and FIMI processing. Although the exploratory EEG study presented in Appendix A remains preliminary, it highlights the potential of such signals to reveal cognitive and affective responses that are not accessible through behavioral data alone. In particular, such signals may provide complementary information to eye-tracking metrics by offering a window into underlying states associated with attentional engagement, cognitive effort, and affective arousal during exposure to misleading content. The multimodal integration of EEG and eye-tracking data therefore represents a promising avenue for future research aiming to better characterize the temporal dynamics of information processing in fake news contexts. Overall, incorporating neurophysiological signals into fake news

research may contribute to a more fine-grained and comprehensive understanding of the mechanisms underlying both susceptibility to and resistance against FIMI.

APPENDIX A

EEG DATA COLLECTION AND EXPERIMENTAL SETUP

This appendix provides a description of the electroencephalography (EEG) data collection protocol used during the experiment. Although EEG data were collected alongside eye-tracking data, they are not analyzed in the scope of this dissertation and are reserved for future work.

A.1. BACKGROUND

EEG is a widely established neurophysiological technique that measures macroscopically the electrical activity generated by synchronized neuronal populations in the brain [153]. Characterized by an exceptionally high temporal resolution, EEG allows researchers to capture rapid, transient cognitive and affective dynamics that occur during human information processing. Within the broader domain of human-computer interaction, these neurophysiological signals are extensively utilized to map complex mental states, including attentional allocation, emotional reactivity, and cognitive workload fluctuations [154]. For instance, systemic shifts in cognitive effort are robustly associated with power spectral density modulations across specific frequency bands, most notably featuring theta synchronization and alpha desynchronization [155], while frontal alpha asymmetry serves as a primary neural correlate for affective evaluation mechanisms and motivational valences [156].

In the specific context of fake news research, EEG has emerged as a critical tool for moving beyond passive, behavioral observation toward capturing the real-time neural markers of credibility evaluation, attentional engagement, and latent cognitive conflict during exposure to deceptive stimuli. Recent empirical advancements by Abdessalem and Frasson [157], and Croce [158] have significantly advanced this paradigm by operationalizing EEG to investigate the specific neural correlates of deception detection and cognitive dissonance. When individuals encounter misleading claims, particularly those that violate internal coherence or deeply held prior beliefs, the resulting cognitive friction manifests as specific electrophysiological signatures indicative of cognitive conflict.

Despite its high temporal granularity, a primary limitation of EEG is its susceptibility to various muscular and environmental artifacts, meaning that raw signals are inherently noisy and demand rigorous preprocessing pipelines. To overcome these limitations and achieve a more granular understanding of human cognition, contemporary frameworks advocate for a multimodal approach that pairs central nervous system tracking with peripheral behavioral metrics. As shown by Abdessalem and Frasson [157], the synchronous co-registration of EEG and eye-tracking modalities effectively bridges the analytical gap between visual attention and internal neurocognitive processing [159]. While eye-tracking isolates precise fixation instances on linguistic triggers (e.g., proper nouns of political figures), the synchronized EEG data simultaneously captures the corresponding cortical responses. This multimodal integration enables an unmediated, high-fidelity characterization of how System 1 heuristics are either accepted implicitly or actively challenged by System 2 analytical verification during the consumption of fake news.

A.2. RELATED WORK

In addition to eye-tracking studies, a growing body of work has explored the use of EEG to investigate cognitive and affective responses to information processing. EEG provides direct measurements of brain activity and has been widely used to assess attention, cognitive workload, and emotional engagement.

Previous research has shown that exposure to misleading or emotionally charged information can elicit distinct neural responses, particularly in frontal brain regions associated with cognitive control and affective evaluation [156]. Variations in EEG frequency bands, such as increased theta activity, have been associated with higher cognitive effort and conflict processing [155].

More recent studies have combined EEG with machine learning techniques to classify cognitive states during information consumption. For example, Chakladar et al. [79] showed that EEG signals can be used to detect emotional responses associated with different types of media content. These

findings suggest that EEG provides complementary insights into the internal processes underlying user interactions with potentially misleading information.

Neuroimaging studies, particularly those employing electroencephalography (EEG), offer valuable insights into the brain's response to news content. For example, EEG studies have revealed that individuals may not differentiate between true and fake news at the neural level, as evidenced by the lack of significant differences in brain activity [160]. Such findings highlight the challenges the brain faces in distinguishing between true and fake news, despite their potential implications for decision-making.

Arisoy et al. [160] investigates whether the human brain is able to distinguish between true and false textual information by measuring neuro-cognitive responses using EEG data. The authors conducted an experiment where 19 participants read both real and fake news headlines while their brain activity was recorded. The study found that there was no significant difference in the participants' EEG responses to true versus fake news, suggesting that, at a neuro-cognitive level, the human brain does not reliably detect textual fake news. The authors conclude that combating fake news requires external interventions such as technological fact-checking tools or literacy training rather than relying solely on individuals' innate cognitive abilities to spot fake news. Our work extends this study by integrating ocular data and the study of not_sure class.

However, compared to eye-tracking, the use of EEG in fake news research remains relatively limited. Most existing studies focus on general cognitive and affective responses rather than identifying the specific textual features that trigger these neural reactions. This limitation highlights the need for multimodal approaches that integrate behavioral and physiological signals to better understand how users process fake news.

A.3. METHODOLOGY

The proposed methodology follows a multi-stage pipeline consisting of experimental setup, data acquisition, configuration, anad metrics.

A.3.1. EXPERIMENTAL SETUP

During the single-session study, participants evaluated 30 news claims presented in randomized order while eye-tracking data, EEG data and participants responses were recorded.

A.3.2. EEG DATA ACQUISITION

In addition to eye-tracking, electroencephalography (EEG) signals were recorded using the Emotiv EPOC+ neuroheadset. The Emotiv EPOC+ is a portable EEG device equipped with 14 electrodes positioned according to the international 10–20 system. It provides a sampling rate of 128 Hz and allows the capture of real-time brain activity related to cognitive and emotional processes.

The EEG system was used to complement eye-tracking data by capturing neural responses associated with attention and cognitive load during the task. The wireless and non-invasive design of the Emotiv EPOC+ makes it suitable for experimental settings involving human-computer interaction.

Both eye-tracking and EEG data were synchronized to enable a multimodal analysis of visual attention and cognitive processing during fake news evaluation.

A.3.3. EEG CONFIGURATION

The Emotiv Epoc+ headset was positioned on each participant following standardized electrode placement protocols. Contact quality for all 14 channels was verified to ensure adequate signal acquisition throughout the session as shown in Figure A.1.

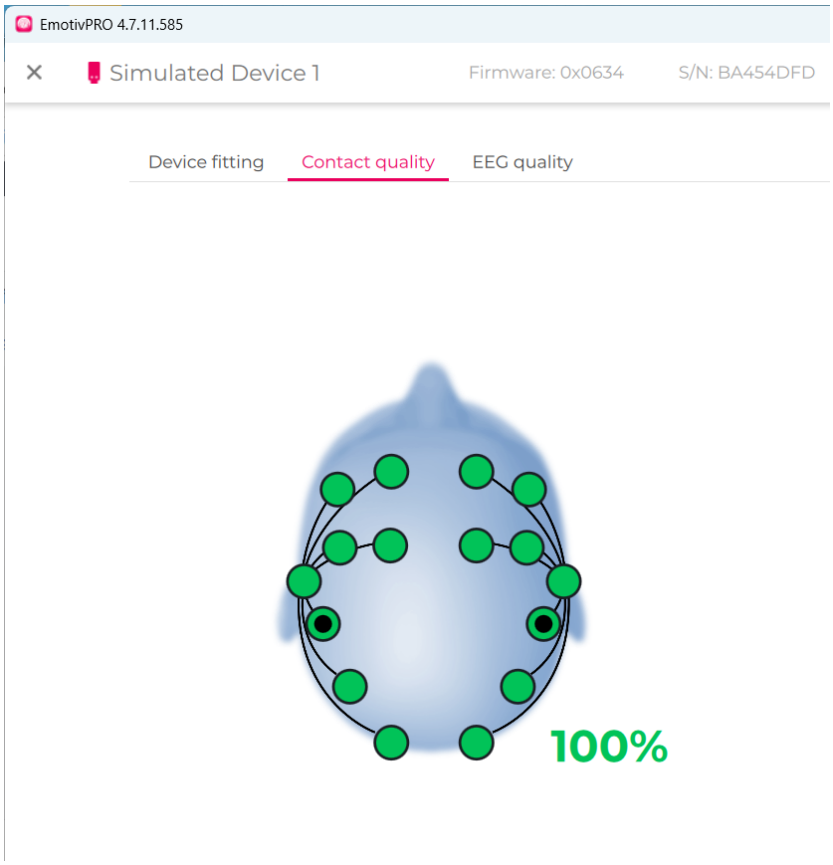


Figure A.1 : Contact quality

A.3.4. EEG METRICS

Electroencephalography (EEG) headsets record electrical activity in the brain across various frequency bands (delta, theta, alpha, beta, gamma), each of which has been empirically linked to distinct mental states such as relaxation, attention, or cognitive workload.

The **Emotiv Epoc+** headset enables the collection of EEG signals. It is designed to provide comprehensive coverage of the frontal and prefrontal regions, as well as the temporal, parietal, and occipital lobes. The use of the physiological sensors in the headset is painless and risk-free. It's a passive device that simply reads the brain's natural electrical activity without sending any electrical signals. Figure A.2 shows the device and its positioning on the head. Access to raw EEG data required a licensed subscription, which includes pretrained models provided by the manufacturer.

Figure A.2 : Emotiv Epoc+ headset

The Emotiv Epoc+ headset provides access to six performance-related EEG metrics: engagement, stress, excitement, relaxation, interest, and focus. The following interpretations were used in the context of our experiment:

- Engagement: Measures active cognitive involvement and mental participation with the content. Higher values indicate that the participant is mentally invested and actively processing the information, while lower values suggest passive consumption or cognitive disengagement from the content.
- Stress : Reflects mental tension (frustration) or cognitive discomfort. Higher values may occur when participants are exposed to disturbing, contradictory, or cognitively demanding news.
- Excitement: Indicates emotional arousal and general stimulation. Elevated values are expected for surprising or emotionally charged content, whether positive or negative.
- Relaxation: Represents a state of mental calmness. Higher relaxation values suggest a peaceful or neutral reading experience, whereas stressful news may reduce this signal.
- Interest: Captures cognitive engagement and affinity toward the content. Values are expected to vary depending on personal relevance or perceived importance of the news for each participant.

- Focus: Measures cognitive concentration. Higher values suggest sustained reading focus, whereas distraction or disengagement results in lower scores.

Figure A.3 illustrates the synchronized data structure for a single participant response. The visualization displays the temporal alignment of EEG-derived emotional metrics captured during exposure to one news claim. Each bar represents the average amplitude of a specific emotional indicator (engagement, stress, relaxation, excitement, interest and focus) recorded throughout the stimulus presentation period. The synchronized timestamps allow for direct correlation between emotional states, eye-tracking patterns, and the participant's responses. This integrated data format facilitates the multimodal analysis by ensuring that all physiological responses are temporally linked to the specific moment of news exposure and cognitive processing.

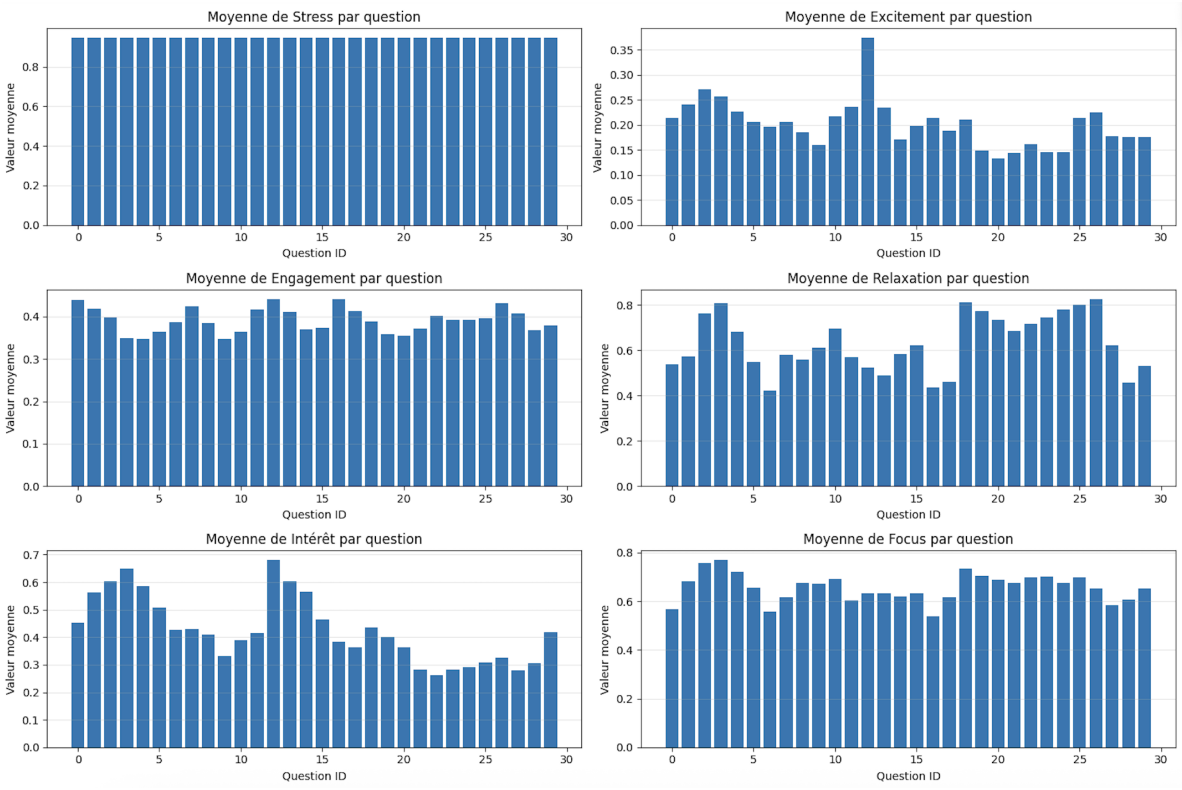


Figure A.3 : Example of EEG average metrics for a participant's response

During each recording session, the system outputs a percentage distribution across these emotional states. For example, a participant may exhibit 60% excitement, 20% relaxation, and 40% stress in response to a given claim.

To ensure the accuracy of the emotional metrics, each measure is accompanied by a validity flag that operates independently. If the signal quality is too poor, the metric will return a null value, indicating the data is unreliable. Additionally, an Overall Quality Score (ranging from 0 to 100) assesses the headset's contact with the scalp. A score above 70 is considered sufficient for reliable data collection across most channels.

A.4. POTENTIAL INTEGRATION OF EEG AND EYE-TRACKING MEASURES

The combined eye-tracking and EEG framework provides a unique opportunity to investigate not only where participants allocate visual attention, but also the potential neurocognitive states associated with these attentional patterns during fake news processing.

The present findings suggest that PPE may function as heuristic credibility cues that reduce analytical scrutiny during information processing. The significant reduction in visit duration, combined with the absence of increased pupillary response, indicates that participants may process PPE-containing fake news more fluently and with reduced cognitive effort.

From a neurophysiological perspective, this pattern could theoretically correspond to lower levels of cognitive conflict and analytical engagement, potentially reflected in EEG-derived indicators such as *reduced engagement or increased relaxation*. Although the EEG signals collected in this experiment were not analyzed within the scope of this dissertation, the synchronized acquisition of eye-tracking and EEG data establishes a promising multimodal framework for future investigations into the neurocognitive mechanisms underlying heuristic processing and misinformation susceptibility.

REFERENCES

[1] Y. Sun, J. He, L. Cui, S. Lei, and C.-T. Lu, "Exploring the deceptive power of LLM-generated fake news: A study of real-world detection challenges," *arXiv preprint arXiv:2403.18249*, 2024.

[2] D. Sallami, Y.-C. Chang, and E. Aïmeur, "From deception to detection: The dual roles of large language models in fake news," *arXiv preprint arXiv:2409.17416*, 2024.

[3] R. Ladouceur, "Cyber influence stakes," in *Blockchain and Artificial Intelligence-Based Solutions to Enhance Privacy in Digital Identity and IoT*, F. Jaafar and S. Pierre, Eds. CRC Press, 2023, pp. 155–174.

[4] R. Ladouceur, S. Pierre, and F. Jaafar, "Cyberwarfare: Geopolitics of technological infrastructures and social networks," in *Major Catastrophes in the 21st Century: Health, Environment, Food, War*, ser. Collection Magna Carta, D. Uzunidis and L. Adatto, Eds. Editions Le Manuscrit, 2022.

[5] European Union Agency for Cybersecurity (ENISA), "ENISA Threat Landscape 2025," European Union Agency for Cybersecurity (ENISA), Report, Oct. 2025, accessed: 2026-01-21. [Online]. Available: https://www.enisa.europa.eu/sites/default/files/2025-10/ENISA%20Threat%20Landscape%202025%20Booklet.pdf

[6] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[7] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. Melbourne, VIC, Australia: Association for Computing Machinery, 2019, pp. 312–320.

[8] World Health Organization. (2019) Ten threats to global health in 2019. World Health Organization. [Online]. Available: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019

[9] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign," in *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*. IEEE, 2018, pp. 258–265.

[10] A. Boily, "Selon Facebook, la Russie est le principal produc-

teur de désinformation," https://www.journaldemontreal.com/2021/05/28/selon-facebook-la-russie-est-le-principal-producteur-de-desinformation, May 2021.

[11] Reuters, "Meta's community notes model will not apply to paid ads," https://www.reuters.com/technology/metas-community-notes-model-will-not-apply-paid-ads-wsj-reports-2025-01-17, Jan. 2025, accessed: 2025-02-06.

[12] European Union Agency for Cybersecurity, "ENISA threat landscape 2022," European Union Agency for Cybersecurity (ENISA), Tech. Rep., 2022, accessed: 2024-04-06. [Online]. Available: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

[13] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun, "User preference-aware fake news detection," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 2051–2055.

[14] M. Szczepański, M. Pawlicki, R. Kozik, and M. Choraś, "New explainability method for BERT-based model in fake news detection," *Scientific reports*, vol. 11, no. 1, p. 23705, 2021.

[15] M. N. S. Roslan and M. Mohd, "Performance comparison of zero-shot and two-shot prompting in detecting fake news using large language models," *Malaysian Journal of Computer Science*, vol. 38, no. Special Issue, 2025.

[16] L. Wu, X. Jiang, S. Sun, Y. Lei, T. Wen, Y. Wang, and M. Liu, "ZoFia: Zero-shot fake news detection with entity-guided retrieval and multi-LLM interaction," *arXiv preprint arXiv:2511.01188*, 2025. [Online]. Available: https://arxiv.org/abs/2511.01188

[17] T. Vergho, J.-F. Godbout, R. Rabbany, and K. Pelrine, "Comparing GPT-4 and open-source language models in misinformation mitigation," *arXiv preprint arXiv:2401.06920*, 2024. [Online]. Available: https://arxiv.org/abs/2401.06920

[18] M. R. Haupt, L. Yang, T. Purnat, and T. Mackey, "Evaluating the influence of role-playing prompts on chatGPT's misinformation detection accuracy: Quantitative study," *JMIR Infodemiology*, vol. 4, p. e60678, 2024.

[19] C. Chen and K. Shu, "Can LLM-generated misinformation be detected?" *arXiv*, vol. 2309.13788, 2023, arXiv preprint, cs.CL, accepted to ICLR 2024.

[20] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[21] G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand, "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management science*, vol. 66, no. 11, pp. 4944–4957, 2020.

[22] Ömer Sümer, E. Bozkir, T. C. Kübler, S. Grüner, S. Utz, and C. Wolff, "Fakenewsperception: An eye movement dataset on the perceived believability of news stories," *Data in Brief*, vol. 36, p. 107044, 2021.

[23] J. Simko, P. Racsko, M. Tomlein, M. Hanakova, R. Moro, and M. Bielikova, "A study of fake news reading and annotating in social media context," *New Review of Hypermedia and Multimedia*, 2021.

[24] L. Shi, N. Bhattacharya, A. Das, and J. Gwizdka, "True or false? cognitive load when reading COVID-19 news headlines: An eye-tracking study," in *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, ser. CHIIR '23. New York, NY, USA: Association for Computing Machinery, mar 2023, pp. 107–116.

[25] S. Dávila-Montero, J. A. Dana-Lê, G. Bente, A. T. Hall, and A. J. Mason, "Review and challenges of technologies for real-time human behavior monitoring," *IEEE transactions on biomedical circuits and systems*, vol. 15, no. 1, pp. 2–28, 2021.

[26] M. Sekicki and M. Staudte, "Eye'll help you out! how the gaze cue reduces the cognitive load required for reference processing," *Cognitive Science*, vol. 42, no. 8, pp. 2418–2458, 2018.

[27] B. M. Wellons and C. N. Wahlheim, "Misinformation reminders enhance belief updating and memory for corrections: The role of attention during encoding revealed by eye tracking," *Cognitive Research: Principles and Implications*, vol. 10, p. 39, 2025.

[28] Sun Tzu, *The Art of War*. London: Luzac & Company, 1910, texte chinois du V^{e} siècle av. J.-C.

[29] W. J. Campbell, *Yellow Journalism: Puncturing the Myths, Defining the Legacies*. Westport, CT: Praeger, 2001.

[30] M. A. Blanchard, "The meaning of the USS maine explosion," *American Journalism*, vol. 8, no. 1, pp. 9–24, 1991.

[31] T. Rid and L. Vikan, "The soviet union's aids disinformation campaign," *Journal of Strategic Studies*, vol. 41, no. 1–2, pp. 131–161, 2018.

[32] A. Henschke, M. Sussex, and C. O'Connor, "Countering foreign interference: election integrity lessons for liberal democracies," *Journal of Cyber Policy*, vol. 5, no. 2, pp. 180–198, 2020.

[33] K. Bleyer-Simon and U. Reviglio, "Defining disinformation across EU and VLOP policies," 2024, consulted February 25, 2026. [Online]. Available: https://edmo.eu/wp-content/uploads/2024/10/EDMO-Report-âĂŞ-Defining-Disinformation-across-EU-and-VLOP-Policies.pdf

[34] Government of Canada, "National cyber threat assessment 2020," Canadian Centre for Cyber Security, Tech. Rep., 2020, retrieved from https://www.cyber.gc.ca/sites/default/files/publications/ncta-2020-e-web.pdf. [Online]. Available: https://www.cyber.gc.ca/sites/default/files/publications/ncta-2020-e-web.pdf

[35] P. J. MacKenzie, "Cyberspace and cyber-enabled information warfare," in *Proceedings of the Joint Air & Space Power Conference 2018*. Essen, Germany: Joint Air Power Competence Centre (JAPCC), 2018. [Online]. Available: https://www.japcc.org/cyberspace-and-cyber-enabled-information-warfare

[36] Government of Canada, "Statement from the minister of national defence: Cyber threats to critical infrastructure," Communications Security Establishment, April 2023, government of Canada official statement. [Online]. Available: https://www.canada.ca/en/communications-security/news/2023/04/statement-from-the-minister-of-national-defence--cyber-threats-to-critical-infrastructure.html

[37] E. Desrosiers. (2023) Le coût économique de la propagation de fausses informations. Le Devoir. [Online]. Available: https://www.ledevoir.com/economie/779860/coronavirus-le-cout-economique-de-la-propagation-de-fausses-informations

[38] D. O. Klein and J. R. Wueller, "Fake news: A legal perspective," *Australasian Policing*, vol. 10, no. 2, pp. 11–17, 2018.

[39] Center for Countering Digital Hate, "Toxic twitter II: Final report," Center for Countering Digital Hate, Tech. Rep., March 2023, accessed July 23, 2023. [Online]. Available: https://counterhate.com/wp-content/uploads/2023/03/Toxic-Twitter-II-Final-Report.pdf

[40] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is

difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[41] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen netzen," Master's thesis, Technische Universität München, Munich, Germany, 1991, advisor: P. Caballero.

[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[43] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[46] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[47] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 8440–8451.

[48] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini *et al.*, "Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference," *arXiv preprint arXiv:2412.13663*, 2024.

[49] S. Jain and B. C. Wallace, "Attention is not explanation," 2019. [Online]. Available: https://arxiv.org/abs/1902.10186

[50] J. Alghamdi, Y. Lin, and S. Luo, "Modeling fake news detection using bert-cnn-bilstm architecture,"

in *2022 IEEE 5th international conference on multimedia information processing and retrieval (MIPR)*, 2022, pp. 354–357.

[51] Uxai, "Explainable AI (XAI)," https://www.uxai.design/explainable-ai, 2024, accessed July 18, 2024.

[52] R. Brown, "Why we need explainable AI," LinkedIn, Jul. 2023. [Online]. Available: https://www.linkedin.com/pulse/why-we-need-explainable-ai-rebel-brown/

[53] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—explainable artificial intelligence," *Communications of the ACM*, vol. 64, no. 6, pp. 44–58, 2021.

[54] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[55] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017, pp. 4765–4774.

[56] T. B. Brown, B. Mann, N. Ryder, and et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, 2020.

[57] J. Kaplan, S. McCandlish, T. Henighan, and et al., "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[58] L. Ouyang, J. Wu, X. Jiang, and et al., "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, 2022.

[59] HumAI Blog. (2026, January) Best AI models 2026: GPT-5 vs Claude 4.5 Opus vs Gemini 3 Pro (complete comparison). Accessed: 2026-03-03. [Online]. Available: https://www.humai.blog/best-ai-models-2026-gpt-5-vs-claude-4-5-opus-vs-gemini-3-pro-complete-comparison/

[60] Macaron. (2025, November) 2025 AI battle: Gemini 3, ChatGPT 5.1 & Claude 4.5. Accessed: 2026-03-03. [Online]. Available: https://macaron.im/blog/gemeni3-vs-gpt-vs-claude

[61] DataStudios. (2025, June) ChatGPT vs Claude vs Gemini: Full report and comparison of features, performance, integrations, pricing, and use cases.

Accessed: 2026-03-03. [Online]. Available: https://www.datastudios.org/post/chatgpt-vs-claude-vs-gemini-full-report-and-comparison-of-features-performance-integrations-pric

[62] J. Ip, "LLM evaluation metrics: Everything you need for LLM evaluation," https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation, May 2026, confident AI Blog.

[63] S. Kula, R. Kozik, and M. Choraś, "Implementation of the BERT-derived architectures to tackle disinformation challenges," *Neural Computing and Applications*, vol. 34, no. 23, pp. 20 449–20 461, 2022.

[64] Y. Liu, Y. Wang, Y. Zhang, X. Chen, and J. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[65] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu *et al.*, "Judging LLMs by LLMs: A benchmark for LLM evaluation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[66] W. Kryściński, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[67] A. Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[68] E. Durmus, H. He, and M. Diab, "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[69] CloudInsight Team, "LLM model ranking & comparison: 2026 major large language model benchmark review," 2026. [Online]. Available: https://cloudinsight.cc/en/blog/llm-ranking

[70] I. Pahima, "Best LLM models in 2026: Complete rankings & comparison," 2026. [Online]. Available: https://www.prompt-compare.ai/blog/best-llm-models-2026

[71] D. Kahneman, *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux, 2011.

[72] C. Unkelbach, "Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 33, no. 1, pp. 219–230, 2007.

[73] B. Lutz, M. T. Adam, S. Feuerriegel, N. Pröllochs, and D. Neumann, "Affective information processing of fake news: Evidence from neurois," *European Journal of Information Systems*, vol. 33, no. 5, pp. 654–673, 2024.

[74] G. Pennycook and D. G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning," *Cognition*, vol. 188, pp. 39–50, 2019.

[75] P. Walla, "Affective processing guides behavior and emotions communicate feelings: Towards a guideline for the NeuroIS community," in *Information Systems and Neuroscience: Gmunden Retreat on NeuroIS 2017*, ser. Lecture Notes in Information Systems and Organisation, R. Riedl, P.-M. Léger, F. D. Davis, and G. R. on NeuroIS, Eds. Cham, Switzerland: Springer International Publishing, 2017, vol. 25, pp. 141–150.

[76] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[77] M. Osmundsen, A. Bor, P. B. Vahlstrup, A. Bechmann, and M. B. Petersen, "Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter," *American Political Science Review*, vol. 115, no. 3, pp. 999–1015, 2021.

[78] K. R. Scherer, A. Schorr, and T. Johnstone, Eds., *Appraisal processes in emotion: theory, methods, research*, ser. Series in Affective Science. New York, NY: Oxford University Press, 2001.

[79] J. Chakladar and R. Chakraborty, "Emotion detection using physiological signals: a survey," *Biomedical Signal Processing and Control*, vol. 39, pp. 101–118, 2018.

[80] S. Horlings, S. Schouten, P. A. M. Kommers, and H. J. J. van Merriënboer, "Effects of slow- and fast-paced game play on motivational and cognitive engagement in a mathematics game," *Educational Technology Research and Development*, vol. 56, no. 5-6, pp. 439–460, 2008.

[81] P. Lallé, B. Berthelot, and D. Fleck, "Eye tracking and emotion recognition in learning contexts," in *Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction*, ser. ACII '17. San Antonio, TX, USA: IEEE, 2017, pp. 123–130. [Online]. Available:

https://doi.org/10.1109/ACII.2017.8273590

[82] D. Kahneman and J. Beatty, "Pupil diameter and load on memory," *Science*, vol. 154, no. 3756, pp. 1583–1585, 1966.

[83] W. Y. Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers)*, 2017, pp. 422–426.

[84] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint*, 2019.

[85] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 626–637.

[86] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model," *Digital Threats: Research and Practice*, vol. 1, no. 2, pp. 1–25, 2020.

[87] K. Shu, G. Zheng, Y. Li, S. Mukherjee, A. H. Awadallah, S. Ruston, and H. Liu, "Early detection of fake news with multi-source weak social supervision," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*. Springer International Publishing, 2021, pp. 650–666.

[88] M. Alizadeh, J. N. Shapiro, C. Buntain, and J. A. Tucker, "Content-based features predict social media influence operations," *Science advances*, vol. 6, no. 30, p. eabb5824, 2020.

[89] S. Alhazbi, "Behavior-based machine learning approaches to identify state-sponsored trolls on Twitter," *IEEE Access*, vol. 8, pp. 195 132–195 141, 2020.

[90] M. Heidari, J. H. Jones, and O. Uzuner, "Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter," in *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, November 2020, pp. 480–487.

[91] S. T. Smith, E. K. Kao, E. D. Mackin, D. C. Shah, O. Simek, and D. B. Rubin, "Automatic detection of influential actors in disinformation networks," *Proceedings of the National Academy of Sciences*, vol. 118, no. 4, p. e2011216118, 2021.

[92] H.-S. Jwa, D. Oh, K. Park, J.-M. Kang, and H. Lim, "exBAKE: Automatic fake news detection model based on BERT," *Applied Sciences*, vol. 9, no. 19, p. 4062, 2019.

[93] M. Heidari, S. Zad, P. Hajibabaee, M. Malekzadeh, and A. Agah, "BERT model for fake news detection based on social bot activities in the COVID-19 pandemic," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2021, pp. 1040–1045.

[94] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia tools and applications*, vol. 80, no. 8, pp. 11 765–11 788, 2021.

[95] S. Kula, M. Choraś, and R. Kozik, "Application of the BERT-based architecture in fake news detection," in *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, vol. 12. Springer International Publishing, 2021, pp. 239–249.

[96] A. Glazkova, M. Glazkov, and T. Trifonov, "g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection," in *Combating Online Hostile Posts in Regional Languages during Emergency Situations: First International Workshop, Constraint 2021, Colocated with AAAI 2021*, 2021, pp. 116–127.

[97] A. J. Keya, M. A. H. Wadud, M. F. Mridha, M. Alatiyyah, and M. A. Hamid, "AugFake-BERT: Handling imbalance through augmentation of fake news using BERT to enhance the performance of fake news classification," *Applied Sciences*, vol. 12, no. 17, p. 8398, 2022.

[98] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 395–405.

[99] S. Mohseni, F. Yang, S. Pentyala, M. Du, Y. Liu, N. Lupfer, X. Hu, S. Ji, and E. Ragan, "Machine learning explanations to prevent overtrust in fake news detection," in *Proceedings of the international AAAI conference on web and social media*, vol. 15, 2021, pp. 421–431.

[100] S.-Y. Chien, C.-J. Yang, and F. Yu, "XFlag: Explainable fake news detection model on social media," *International Journal of Human–Computer Interaction*, vol. 38, no. 18-20, pp. 1808–1827, 2022.

[101] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information*

Processing Systems (NeurIPS), vol. 35, pp. 24 824–24 837, 2022.

[102] S. Krishna, J. Ma, D. Slack, A. Ghandeharioun, S. Singh, and H. Lakkaraju, "Post hoc explanations of language models can improve language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[103] M. Gao *et al.*, "LLM-based NLG evaluation: Current status and challenges," *arXiv preprint arXiv:2402.01383*, 2024.

[104] Y. Wang, Z. Gu, S. Zhang, S. Zheng, T. Wang, T. Li, H. Feng, and Y. Xiao, "LLM-GAN: Construct generative adversarial network through large language models for explainable fake news detection," *arXiv preprint arXiv:2409.01787*, 2024.

[105] K. Kim, S. Lee, K.-H. Huang, H. P. Chan, M. Li, and H. Ji, "Can LLMs produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate," *arXiv preprint arXiv:2402.07401*, February 2024.

[106] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, and Z. Xu, "Deepseek-V2: A strong, economical, and efficient mixture-of-experts language model," *arXiv preprint arXiv:2405.04434*, 2024.

[107] M. Boissonneault, G. Bouchard, and F. Gagnon, "Fake news detection: Evaluating the performance of ChatGPT," *arXiv preprint arXiv:2401.12345*, 2024.

[108] E. Bozkir, G. Kasneci, S. Utz, and E. Kasneci, "Regressive saccadic eye movements on fake news," in *2022 Symposium on Eye Tracking Research and Applications*. New York, NY, USA: Association for Computing Machinery, 2022, pp. 1–7.

[109] N. Steinfeld, "How do users examine online messages to determine if they are credible? an eye-tracking study of digital literacy, visual attention to metadata, and success in misinformation identification," *Social Media + Society*, vol. 9, no. 3, p. 20563051231196871, aug 2023.

[110] E. D. Filippi, A. Nandi, and A. P. Baños, "Misinformation detection through multimodal machine learning analysis of eye-tracking and physiological responses: A proof-of-concept study," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 18–33.

[111] B. Lutz, M. Adam, S. Feuerriegel, N. Pröllochs, and D. Neumann, "Which linguistic cues make

people fall for fake news? a comparison of cognitive and affective processing," *Proceedings of the ACM on human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–22, 2024.

[112] R. Ladouceur, A. Hassan, M. Mejri, and F. Jaafar, "Can deep learning detect fake news better when adding context features?" in *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 2024, pp. 56–61.

[113] S. Raza and C. Ding, "Fake news detection based on news content and social contexts: a transformer-based approach," *International Journal of Data Science and Analytics*, vol. 13, no. 4, pp. 335–362, 2022.

[114] D. A. Martin and J. N. Shapiro, "Trends in online foreign influence efforts," Princeton University, Princeton, NJ, Working Paper, 2019.

[115] Centre canadien pour la cybersécurité, "Évaluation des cybermenaces nationales 2025-2026," Centre de la sécurité des télécommunications, Ottawa, ON, Rapport, Oct. 2024, consulté le 21 janvier 2026. [Online]. Available: https://www.cyber.gc.ca/fr/orientation/evaluation-cybermenaces-nationales-2025-2026

[116] M.-J. Hogue, "Enquête publique sur l'ingérence étrangère dans les processus électoraux et les institutions démocratiques fédéraux : rapport final," Commission sur l'ingérence étrangère, Gouvernement du Canada, Ottawa, ON, Tech. Rep., janvier 2025, rapport en 7 volumes. ISBN 978-0-660-75086-6. [Online]. Available: https://commissioningerenceetrangere.ca/

[117] C. Cunningham, *Cyber Warfare–Truth, Tactics, and Strategies: Strategic concepts and truths to help you and your organization survive on the battleground of cyber warfare*. Packt Publishing Ltd, 2020.

[118] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, May 2017, pp. 280–289.

[119] L. Asmelash. (2020, July) How black lives matter went from a hashtag to a global rallying cry. CNN. [Online]. Available: https://www.cnn.com/2020/07/26/us/black-lives-matter-explainer-trnd/index.html

[120] S. Aral, *The Hype Machine: How social media disrupts our elections, our economy, and our health*

and how we must adapt. New York: Currency, 2021.

[121] P. Lapointe. (2022, Feb.) Contrer la désinformation au québec: une lutte inégale contre les géants du web. Retrieved from J-Source. [Online]. Available: https://j-source.ca/contrer-la-desinformation-au-quebec-une-lutte-inegale-contre-les-geants-du-web/

[122] European Commission. (2024) Digital services act. European Commission. Consulté le 28 juin 2026. [Online]. Available: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

[123] R. Negi, A. Walia, D. Singh, D. Garg, and P. S. Rana, "Enhancing fake news detection: The critical role of model fine-tuning for optimal performance," in *Proceedings of the International Conference on Information Technology and Artificial Intelligence (ITAI 2025)*, ser. Lecture Notes in Networks and Systems. Springer, 2025, online First. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-96-8694-0_11

[124] X. Zhou, R. Zafarani, and E. Ferrara, "From fake news to# fakenews: Mining direct and indirect relationships among hashtags for fake news detection," *arXiv preprint arXiv:2211.11113*, 2022.

[125] R. Ladouceur, C. Njeh, H. Nakouri, and F. Jaafar, "Comparisons between open-LLMs for fake news detection," in *2025 9th Cyber Security in Networking Conference (CSNet)*. IEEE, 2025, pp. 1–5.

[126] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 65–72.

[127] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020.

[128] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for mt evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 2685–2702.

[129] K. Pelrine, A. Imouza, C. Thibault, M. Reksoprodjo, C. A. Gupta, J. N. Christoph, J.-F. Godbout, and R. Rabbany, "Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4," *arXiv preprint arXiv:2305.14928*, 2023.

[130] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung, "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," in *Proceedings of IJCNLP-AACL*, 2023.

[131] O. Buchholz, "A means-end account of explainable artificial intelligence," *Synthese*, 2023.

[132] Y. Huang, K. Shu, P. S. Yu, and L. Sun, "FakeGPT: Fake news generation, explanation and detection of large language models," *arXiv preprint arXiv:2310.05046*, 2023.

[133] M. Gao, X. Hu, X. Yin, J. Ruan, X. Pu, and X. Wan, "LLM-based nlg evaluation: Current status and challenges," *Computational Linguistics*, pp. 1–27, 2025.

[134] A. Rangapur, H. Wang, and K. Shu, "Investigating online financial misinformation and its consequences: A computational perspective," *arXiv preprint arXiv:2309.12363*, 2023.

[135] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 conference on empirical methods in natural language processing*, 2023, pp. 12 076–12 100.

[136] R. Ladouceur, T. Pierre, F. Jaafar, and H. Ben Abdessalem, "Elements of credibility or credulity about fake news exposure: An empirical study," 2026, submitted to ACM Transactions on Social Computing, April 15, 2026.

[137] C. Deppe and G. S. Schaal, "Cognitive warfare: a conceptual analysis of the NATO ACT cognitive warfare exploratory concept," *Frontiers in Big Data*, vol. 7, 2024, accessed: 2026-03-04. [Online]. Available: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1452129/full

[138] R. E. Petty and J. T. Cacioppo, "The elaboration likelihood model of persuasion," *Advances in Experimental Social Psychology*, vol. 19, pp. 123–205, 1986.

[139] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Lawrence Erlbaum Associates, 1988.

[140] A. Field, *Discovering statistics using IBM SPSS statistics*, 4th ed. London, England: Sage Publications, 2013.

[141] B. G. Tabachnick and L. S. Fidell, *Using multivariate statistics*, 7th ed. Pearson, 2019.

[142] C. M. Judd, J. Westfall, and D. A. Kenny, "Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem." *Journal of personality and social psychology*, vol. 103, no. 1, p. 54, 2012.

[143] S. Chaiken, "Heuristic versus systematic information processing and the use of source versus message cues in persuasion," *Journal of Personality and Social Psychology*, vol. 39, no. 5, pp. 752–766, 1980.

[144] Cision. (2025) Analyse de 5 années de désinformation : Un réel impact sur la société. [Online]. Available: https://www.cision.fr/nous-connaitre/communiques-de-presse/analyse-de-5-annees-de-desinformation/

[145] G. Pennycook and D. G. Rand, "The psychology of fake news," *Trends in cognitive sciences*, vol. 25, no. 5, pp. 388–402, 2021.

[146] D. Carrington. (2025, jun) Climate misinformation turning crisis into catastrophe: IPIE report. The Guardian. [Online]. Available: https://www.theguardian.com/environment/2025/jun/19/climate-misinformation-turning-crisis-into-catastrophe-ipie-report

[147] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of General Psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[148] NATO Allied Command Transformation, "Cognitive warfare," https://www.act.nato.int/activities/cognitive-warfare/, 2026, accessed: 2026-03-04.

[149] J. G. Adair, "The hawthorne effect: A reconsideration of the methodological artifact," *Journal of Applied Psychology*, vol. 69, no. 2, pp. 334–345, 1984.

[150] A. Shupta, P. Radiuk, M. Kvassay, and P. Gaj, "Verifiable by construction: Evidence-anchored llms for explainable fake news detection." in *ExplAI*, 2025, pp. 168–182.

[151] B. Huang and K. M. Carley, "Disinformation and misinformation on twitter during the novel coronavirus outbreak," *arXiv preprint arXiv:2006.04278*, 2020. [Online]. Available: https://arxiv.org/abs/2006.04278

[152] R. Marigliano, J. Shaha, and K. M. Carley, "Expanding the BEND framework: Enhancing influence prediction in social media networks with advanced metrics," in *Proceedings of the 2024 IDeaS Conference on Disinformation, Hate Speech, and Extremism Online*. Center for Informed Democracy & Social-cybersecurity (IDeaS), Carnegie Mellon University, 2024.

[153] E. Niedermeyer and F. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins, 2005.

[154] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in *Engineering in Medicine and Biology Society*, 2006, pp. 1355–1358.

[155] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance," *Brain Research Reviews*, vol. 29, no. 2-3, pp. 169–195, 1999.

[156] R. J. Davidson, "What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research," *Biological Psychology*, vol. 67, no. 1-2, pp. 219–233, 2004.

[157] H. Ben Abdessalem, "Système intelligent pour le suivi et l'optimisation de l'état cognitif," Ph.D. dissertation, [University name needed], 2021.

[158] K. A. Croce, "Approaching the intersection of neuroscience and education: merging perspectives to strengthen study design," *SN Social Sciences*, vol. 3, no. 1, p. 1, 2022.

[159] S. Lallé and C. Conati, "Do we need to rethink eye-tracking data interpretation for HCI?" *Proceedings of the 2017 CHI Conference*, 2017.

[160] C. Arisoy, A. Mandal, and N. Saxena, "Human brains can't detect fake news: A neuro-cognitive study of textual disinformation susceptibility," in *2022 19th Annual International Conference on Privacy, Security & Trust (PST)*. IEEE, 2022, pp. 1–12.

ETHICAL CERTIFICATION

Ethical Certification Number : CER-UQAC-2025-1925

Project : Étude des émotions face à la désinformation

UQAC
Université du Québec
à Chicoutimi

Le 27 novembre 2024

À l'attention de :
RACHEL LADOUCEUR

Direction de recherche : Fehmi Jaafar

Titre : Étude des émotions face à la désinformation
N° de référence : 2025-1925

Objet : Approbation éthique de votre projet de recherche

Bonjour,

Votre projet de recherche a fait l'objet d'une évaluation en matière d'éthique de la recherche avec des êtres humains par le Comité d'éthique de la recherche de l'Université du Québec à Chicoutimi (CER-UQAC). Un certificat d'approbation éthique qui atteste de la conformité de votre projet de recherche à la Politique d'éthique de la recherche avec des êtres humains de l'UQAC est émis en date du 27 novembre 2024.

Prenez note que ce certificat est valide jusqu'au **27 novembre 2025**.

Les documents ont été validés et officialisés. Ces documents ont été déposés dans votre projet (voir les documents précédés d'un **carré vert** dans la section «Fichiers»). Ces nouvelles versions sont celles autorisées par le CER-UQAC et devront être utilisées pour votre projet.

Selon la Politique d'éthique de la recherche avec des êtres humains, il est de la responsabilité des chercheurs d'assurer que leurs projets de recherche conservent une approbation éthique pour toute la durée des travaux de recherche et d'informer le CER-UQAC de la fin de ceux-ci. Vous devrez donc obtenir le renouvellement de votre approbation éthique avant l'expiration de ce certificat. La poursuite de la **cueillette de données** auprès des participants, sans certification éthique valide, ou le fait d' **apporter une modification significative** (à la population ciblée, au formulaire de consentement, au protocole d'expérimentation, à la méthode de collecte ou de traitement des données, etc.) **ou affectant le niveau de risque du projet** sans approbation du CER-UQAC représentent des situations relevant de la Politique relative à la conduite responsable en recherche et en création. De plus, vous devez signaler tout incident grave dès qu'il survient ainsi que les modifications apportées à votre projet.

Enfin, nous vous demandons d'utiliser le numéro de référence suivant **2025-1925** pour toute correspondance avec le CER-UQAC.

En vous souhaitant le meilleur des succès dans la réalisation de votre recherche, veuillez recevoir nos salutations distinguées.

Le CER-UQAC

UQAC
Université du Québec
à Chicoutimi

Le 21 novembre 2025

RENOUVELLEMENT DE L'APPROBATION ÉTHIQUE

La présente atteste que le projet de recherche décrit ci-dessous a fait l'objet d'un renouvellement administratif de l'approbation éthique émise par le CER-UQAC pour la raison suivante:

N^{o} de référence : 2025-1925

Titre du projet de recherche : Étude des émotions face à la désinformation

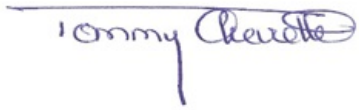
Chercheur principal à l'UQAC
RACHEL LADOUCEUR, Département d'informatique et de mathématique,

Direction / **Codirection de recherche**

En provenance de l'UQAC : Fehmi Jaafar

Date de l'approbation éthique initiale du projet : 27 novembre 2024
Date du prochain renouvellement : 21 novembre 2026
N.B. Un rappel automatique vous sera envoyé par courriel quelques semaines avant l'échéance de votre certificat afin de remplir le formulaire F7 - Renouvellement annuel.

- Si votre projet se termine avant la date du prochain renouvellement, vous devrez remplir le formulaire ***F9 - Fin de projet***.
- Si des modifications sont apportées à votre projet avant l'échéance du certificat, vous devrez remplir le formulaire ***F8 - Modification de projet***.
- Tout nouveau membre de votre équipe de recherche devra être déclaré au CER-UQAC lors de votre prochaine demande de renouvellement ou lors de la fin de votre projet si le renouvellement n'est pas requis. ATTENTION: Vous devez faire signer une déclaration d'honneur aux personnes ayant accès aux participants (ou à des données nominatives sur les participants) et la conserver dans vos dossiers de recherche.
- Si vous avez des cochercheurs dans d'autres universités, veuillez leur transmettre ce certificat.



Tommy Chevrette

Signé le 2025-11-21 à 08:25