UQAC
Université du Québec
à Chicoutimi

AMÉLIORATION DE L'ASSOCIATION ENTRE AVIS D'INTERRUPTION ET INTERRUPTIONS PLANIFIÉES CHEZ HYDRO-QUÉBEC : UNE APPROCHE HYBRIDE COMBINANT APPRENTISSAGE AUTOMATIQUE, IA EXPLICABLE ET MODÈLES DE LANGAGE POUR LE RAFFINEMENT DES RÈGLES D'AFFAIRES

PAR ANGE VALERE WOBINWO AZALY

MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI EN VUE DE L'OBTENTION DU GRADE DE MAÎTRISE ÈS SCIENCES (M. SC.) EN INFORMATIQUE

QUÉBEC, CANADA

REMERCIEMENTS

La réalisation de ce mémoire n'aurait pas été possible sans le soutien et la contribution de nombreuses personnes et institutions, à qui je tiens à exprimer toute ma gratitude.

Je souhaite adresser mes premiers remerciements à mon directeur de recherche, Professeur Kevin Bouchard, dont l'encadrement rigoureux, les conseils avisés et la disponibilité constante ont été des piliers essentiels tout au long de ce projet. Vos orientations éclairées et votre confiance en mes capacités m'ont permis de progresser et de mener à bien ce travail.

Je remercie également chaleureusement mon codirecteur, Professeur Bruno Bouchard, pour ses précieuses contributions, sa lecture attentive de mes travaux et les échanges enrichissants que nous avons eus. Votre expertise complémentaire a grandement contribué à la qualité de ce mémoire.

Mes remerciements vont aussi à mes collègues d'Hydro-Québec pour leur accueil, leur soutien technique et leur générosité dans le partage de leurs connaissances. Votre expertise du terrain et vos retours ont été d'une aide inestimable dans la concrétisation de ce projet.

Je tiens à exprimer ma reconnaissance à Hydro-Québec, l'entreprise au sein de laquelle ce projet a été mené, pour avoir offert un cadre professionnel stimulant et les ressources nécessaires à la réalisation de cette recherche. Cette collaboration a été une expérience formatrice et enrichissante.

Je n'oublie pas de remercier ma famille pour son soutien indéfectible, sa patience et ses encouragements tout au long de ce parcours. Votre présence bienveillante et votre soutien moral ont été une source de motivation constante dans les moments les plus exigeants.

Enfin, je remercie sincèrement tous mes proches qui, de près ou de loin, ont contribué à l'aboutissement de ce mémoire. Que ce soit par vos encouragements, vos relectures ou simplement votre présence réconfortante, votre soutien a été précieux et je vous en suis profondément reconnaissant.

RÉSUMÉ

Chez Hydro-Québec, la gestion du réseau de distribution électrique exige une association fiable entre les avis clients et les interruptions planifiées, condition sine qua non d'un calcul précis de l'indicateur TRIP. L'approche actuelle basée sur des règles déterministes montrent cependant des limites face à la complexité et à l'ambiguïté des situations qui peuvent survenir sur le terrain, générant un volume non négligeable d'entités non associées qui dégradent la fiabilité de cet indicateur.

Ce mémoire propose un cadre exploratoire hybride combinant le Machine Learning classique et les capacités de raisonnement des Large Language Models (LLM). La tâche d'association entre avis et interruptions planifiées est formalisée comme un problème de classification, entraîné sur un jeu de données constitué des associations historiques réalisées à partir des règles d'association de base. Les caractéristiques mobilisées incluent notamment les dates des événements, les causes signalées, les équipements impactés, etc. Les données étant de nature sensible, elles font l'objet de restrictions de publication sans que cela n'affecte la validité des résultats.

Afin de pallier le manque de transparence des modèles de classification dans des contextes décisionnels critiques, le cadre intègre la méthode SHAP (SHapley Additive exPlanations) pour interpréter les prédictions, en ciblant spécifiquement les faux positifs. Les valeurs SHAP, combinées à des échantillons de données, sont ensuite injectées dans un prompt LLM jouant le rôle d'interprète. L'usage de modèles de raisonnement intégrant nativement la technique Chain-of-Thought (CoT) permet de générer des étapes de raisonnement intermédiaires, facilitant l'identification de nouvelles tendances et la généralisation de règles d'association plus robustes.

Les résultats obtenus démontrent que cette approche hybride réduit significativement le nombre d'avis et d'interruptions planifiées non appariés et améliore la fiabilité du calcul de l'indicateur TRIP, tout en offrant une explicabilité opérationnelle que seuls les experts métier sont en mesure d'apprécier.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

BERT Bidirectional Encoder Representations from Transformers
BLEU Bilingual Evaluation Understudy
BPM Business Process Management
BPMN Business Process Model and Notation
CED Centre d'Exploitation et de Distribution
CER-UQAC Comité d'éthique de la recherche de l'Université du Québec à Chicoutimi
CoT Chain-of-Thought
DBML Database Markup Language
DL Deep Learning
EM Entity Matching
EPD Electric Power Domain
GEM Generalized Entity Matching
IA Intelligence Artificielle
KPI Key Performance Indicator
LLM Large Language Model
ML Machine Learning
NLG Natural Language Generation
NLP Natural Language Processing
OMG Object Management Group
RAP Reasoning viA Planning
ReAct Reasoning + Acting
RMSE Root Mean Square Error
RNN Recurrent Neural Network
ROUGE Recall-Oriented Understudy for Gisting Evaluation
RPA Robotic Process Automation
SHAP SHapley Additive exPlanations
SVM Support Vector Machine

TALN Traitement Automatique du Langage Naturel
TRIP Taux de Respect des Interruptions Plannifiées
XAI eXplainable Artificial Intelligence

DÉCLARATION D'UTILISATION DU NIVEAU DE L'INTELLIGENCE ARTIFICIELLE GÉNÉRATIVE (IAG)

L'IA générative a été utilisée pour des suggestions ou des corrections mineures (orthographe, grammaire) ainsi que pour l'amélioration d'éléments visuels, tels que l'ajout d'un titre ou d'une légende. Niveau d'utilisation de l'intelligence artificielle.

CHAPITRE I

INTRODUCTION

1.1 CONTEXTE

À une époque caractérisée par une évolution technologique rapide et une prolifération des données, les organisations doivent continuellement revoir et affiner leurs processus métier afin de maintenir l'excellence opérationnelle et de relever les nouveaux défis [51]. C'est dans ce sens que Dumas et al. [16] définissent la gestion des processus métier, ou *Business Process Management (BPM)* en anglais, comme étant l'art et la science qui consistent à superviser la manière dont le travail est effectué au sein d'une organisation afin de garantir des résultats cohérents et de tirer parti des possibilités d'amélioration [16]. Plus formellement, elle sert de discipline fondamentale agissant comme une boîte à outils complète pour comprendre, documenter, analyser et optimiser des opérations métier complexes [30, 22].

Pour illustrer cette philosophie de gestion de bout en bout, prenons l'exemple d'un processus de demande de congé dans une organisation. Dans une situation non structurée, l'employé peut formuler sa demande de manière informelle, par courriel ou verbalement, ce qui peut entraîner des oublis, des délais de réponse variables ou des incohérences dans l'application des règles internes. En appliquant une démarche de *BPM*, ce processus est d'abord analysé et formalisé.

Il est ensuite modélisé sous forme de workflow clair, comprenant des étapes définies telles que la soumission de la demande par l'employé, la validation par le gestionnaire, la mise à jour du système de paie et la notification automatique de la décision. Ce processus peut alors être partiellement ou entièrement automatisé à l'aide d'un système informatique, assurant une traçabilité des actions, une réduction des erreurs et un traitement plus rapide et uniforme des demandes. Ce cas montre comment l'application du *BPM* permet à une organisation de cartographier et de gérer de manière explicite les étapes nécessaires, les ressources impliquées et les informations échangées du début à la fin, garantissant ainsi que le processus apporte une valeur ajoutée constante.

Alors que le BPM traditionnel se concentre principalement sur la perspective de la cartographie de la séquence des activités et des dépendances, les approches modernes doivent de plus en plus intégrer les perspectives des données pour gérer la complexité de la prise de décision dans le monde réel [30, 28].

L'intégration de l'Intelligence Artificielle (IA) dans le BPM a catalysé un changement de paradigme, réduisant la dépendance à l'égard des efforts manuels et des connaissances spécialisées [19]. Au cœur de cette transformation, les grands modèles de langage, ou Large Language Models (LLMs), se sont imposés comme une rupture majeure : entraînés sur d'énormes corpus de données généralistes, ces algorithmes de nature générative sont capables de comprendre du texte et de produire du nouveau contenu multimodal, servant d'ailleurs de moteur principal derrière les agents conversationnels (*chatbots*) modernes [37]. Appliqués à la gestion des processus, les LLMs s'avèrent des outils puissants pour traiter des entrées textuelles complexes, résoudre des problèmes et générer des sorties structurées. Leurs progrès récents démontrent qu'ils peuvent désormais automatiser la génération de modèles de processus à

partir de descriptions en langage naturel, évaluer l'adéquation des tâches pour l'automatisation robotisée des processus (en anglais *Robotic Process Automation (RPA)*) et prendre en charge les requêtes conversationnelles relatives aux processus [29]. Ces capacités suggèrent que les LLMs ne sont pas de simples processeurs passifs de texte, mais peuvent servir d'agents actifs dans le raffinement et l'exécution des processus métier [30].

1.2 PROBLÉMATIQUE

Malgré ces progrès, il subsiste un écart de recherche notable concernant l'application de l'IA générative dans le cycle de vie du BPM. La littérature actuelle se concentre fortement sur la modélisation des processus ; l'extraction de diagrammes visuels (par exemple, *BPMN*) à partir de descriptions textuelles ; plutôt que sur l'optimisation du processus lui-même [4]. Bien qu'il existe des approches pour la vérification de la conformité, les travaux sur l'utilisation des LLMs pour améliorer activement la logique au sein d'un processus établi, en particulier grâce à l'intégration de techniques d'*eXplainable Artificial Intelligence (XAI)*, sont limités.

Ce vide est particulièrement évident dans le domaine de l'énergie électrique, en anglais *Electric Power Domain (EPD)* [26]. L'intégration des *LLMs* dans ce secteur a été source de transformations. Cependant, la recherche s'est principalement concentrée sur des domaines techniques spécifiques : la prévision de la consommation, le diagnostic des pannes et l'assistance intelligente aux utilisateurs via des chatbots. Si des études ont exploré l'utilisation des *LLMs* pour naviguer dans les réglementations de sécurité, générer des descriptions d'alarmes pour les éoliennes [6], ou encore assurer la gouvernance des données [31], l'application des *LLMs* à l'optimisation des processus métier, tels que l'association des avis-clients avec les interruptions de courant planifiées, reste sous-explorée.

La problématique de ce mémoire s'inscrit dans le contexte de la distribution d'électricité, où la qualité du suivi des interruptions planifiées repose sur la capacité à associer correctement deux types d'entités distinctes : les avis d'interruption communiqués aux clients et les interruptions réellement effectuées sur le réseau. Cette association est essentielle au calcul d'indicateurs de performance stratégiques, tels que le Taux de Respect des Interruptions Planifiées (TRIP), qui influencent à la fois les décisions opérationnelles, l'allocation des ressources et l'évaluation du personnel. Toutefois, en raison de contraintes structurelles des systèmes opérationnels, notamment l'absence de données clients complètes, cette correspondance ne peut être réalisée directement à la source et est plutôt effectuée a posteriori dans un entrepôt de données à l'aide de règles déterministes conçues par des experts. Bien que ces règles permettent de traiter les cas standards, elles montrent des limites face à des situations complexes ou ambiguës, telles que des chevauchements temporels, des incohérences dans les volumes de clients affectés ou des interventions multiples dans une même zone. Ces limites peuvent entraîner des associations manquées, compromettant ainsi la fiabilité des indicateurs produits. Dès lors, la question centrale qui se pose est de savoir dans quelle mesure des approches basées sur l'intelligence artificielle, combinant Machine Learning (ML) classique, intelligence artificielle explicable et modèles de langage, peuvent être mobilisées pour enrichir et améliorer ces règles d'association, tout en conservant leur interprétabilité et leur compatibilité avec les contraintes opérationnelles existantes.

1.3 CONTRIBUTION

Ce mémoire explore un moyen d'amélioration de l'association des avis d'interruptions aux interruptions planifiées, processus opérationnel essentiel au sein d'Hydro-Québec, l'entreprise

en charge de la production, du transport et de la distribution de l'énergie électrique au Québec. Ce processus est basé sur un ensemble de règles formelles et explicites qui déterminent comment lier une interruption planifiée déjà réalisée à un avis précédemment produit. Ces règles prennent généralement en compte les paramètres tels que : le Centre d'Exploitation et de Distribution (CED), les appareils, la plage horaire, etc. Elles visent à assurer que chaque interruption planifiée soit associée à son avis correspondant de manière cohérente. L'objectif final de l'association est le calcul d'un *Key Performance Indicator (KPI)* appelé TRIP.

Cette association semble simple mais est bien plus compliquée dans les faits. Une interruption programmée peut débuter plus tôt et nécessiter le délestage d'une plus grande zone afin de préparer les travaux dans la zone initialement prévue. Ainsi, on se retrouve avec trois (3) interruptions successives de type grande zone - petite zone - grande zone, impactant plus de clients que prévu incluant ceux non avisés.

Outre le scénario précédent, une interruption peut commencer plus tard que prévu pour de nombreuses raisons : retard ou indisponibilité du personnel, trafic sur la route, végétation à abattre, intempéries etc. Toutes ces situations particulières qui peuvent survenir dans la réalité ne permettent pas de faire l'association directement depuis le logiciel source des six (6) CEDs mais plus tard dans un entrepôt de données où l'on retrouvera les données des clients. Quoique malgré l'application des règles, il existe toujours des avis sans interruptions et des interruptions planifiées non rattachées à des avis.

Il devient donc nécessaire de maximiser le nombre d'associations afin fiabiliser le calcul du TRIP. Cet indicateur constitue un indicateur de performance utilisé dans la gestion opérationnelle de l'organisation dans la mesure où il oriente la prise de décision pour une allocation équitable des ressources et une responsabilité claire.

1.4 MÉTHODOLOGIE

Pour résoudre ce problème, cette recherche propose un nouveau cadre qui combine le ML classique et les capacités de raisonnement des LLMs. La méthodologie traite la correspondance d'entités (EM) comme une tâche de classification, en entrainant des algorithmes de classification sur un ensemble de données de paires d'avis et d'interruptions planifiées. Les données mobilisées, de nature sensible, font l'objet de restrictions de publication dont les modalités sont détaillées dans la section méthodologique, sans que cela n'affecte la validité des résultats présentés.

Toutefois,les modèles de classification souffrent fréquemment d'un manque de transparence, ce qui freine leur adoption dans des contextes décisionnels critiques [48, 42, 7]. Pour lever ce verrou, le cadre applique la méthode SHapley Additive exPlanations (SHAP) pour interpréter les prédictions du modèle, en ciblant spécifiquement les faux positifs obtenus lors de l'inférence. SHAP attribue une valeur d'importance à chaque caractéristique [34], quantifiant sa contribution à la prédiction d'une possible association entre un avis et une interruption planifiée.

Considérons par exemple un système d'appariement entre avis de coupure et interruptions planifiées. Un classificateur reçoit en entrée des caractéristiques telles que les dates des événements, les causes ou les appareils impactés. Lorsqu'une association est prédite à tort, l'absence d'explicabilité rend la logique du modèle opaque. L'usage de la méthode SHAP s'avère alors crucial : en décomposant les contributions de chaque variable à la prédiction erronée, elle transforme une erreur de classification en une opportunité d'extraire de nouveaux critères et de généraliser des règles d'association plus robustes.

Cette approche introduit une étape cruciale : l'intégration des résultats SHAP et des échantillons de données dans un *prompt* LLM. En tenant le rôle d'interprète, le LLM fournit un processus amélioré en langage naturel qui aide les experts à trouver de nouvelles tendances. De plus, l'utilisation des modèles dits de raisonnement, qui intègrent nativement la technique *Chain-of-Thought (CoT)* [20], permet de générer des étapes de raisonnement intermédiaires, facilitant ainsi efficacement une logique critique. Cette approche va au-delà de la simple automatisation grâce aux LLMs pour raisonner sur les ambiguïtés des données et améliorer la fiabilité globale du processus.

1.5 ORGANISATION DU MÉMOIRE

Ce document est organisé en six chapitres interconnectés. Le premier chapitre constitue l'introduction générale de ce travail, où sont présentés le contexte de la recherche, la problématique, la contribution du mémoire ainsi que le plan de travail adopté. Ensuite, le deuxième chapitre est consacré au cadre théorique de notre recherche ; il passe en revue les concepts fondamentaux de la gestion des processus métier, de la mise en correspondance d'entités et de l'architecture des LLMs, y compris les mécanismes *Transformer*. Le troisième chapitre fait quant à lui office de revue de littérature afin d'examiner le paysage actuel de l'IA dans le secteur électrique (EPD), les méthodologies existantes pour l'association d'entités (Entity Matching (EM)) et l'utilisation des données pour générer du texte, tout en soulignant les limites des approches actuelles.

Le cœur de notre apport est détaillé dans le quatrième chapitre, qui expose l'hypothèse de recherche et la méthodologie. Plus spécifiquement, il décrit la construction du jeu de données, la conception du classificateur, l'intégration de SHAP et les stratégies de *prompt engineering*

utilisées pour générer de nouvelles règles à partir des LLMs. Par la suite, le cinquième chapitre est dédié aux expérimentations et aux résultats de l'évaluation empirique effectuée pour valider l'approche proposée. Nous y abordons notamment l'usage des métriques automatiques comme BERTScore, l'implication des évaluateurs humains et la notion de *LLM-as-judge*. Enfin, le sixième chapitre conclut ce mémoire autour d'une discussion sur les résultats, les forces et les limitations de l'approche proposée, avant de s'ouvrir sur les orientations futures pour l'optimisation des processus à partir des LLMs.

CHAPITRE II

CADRE THÉORIQUE

Ce chapitre fournit les bases théoriques nécessaires à la compréhension du mémoire. Il comprend également certaines notions connexes qui ne sont pas directement liées au travail détaillé dans le reste du contenu, mais que nous avons jugées pertinentes pour d'éventuels travaux futurs. Il passe en revue les principes fondamentaux du Business Process Management (BPM) et du Entity Matching (EM), avant d'analyser les fondements architecturaux des Large Language Model (LLMs), en particulier l'architecture *Transformer*, qui sert de catalyseur technologique aux avancées modernes dans ces domaines.

2.1 GESTION DES PROCESSUS MÉTIER : DU CYCLE DE VIE À L'AUTOMATISATION

Selon Dumas et al. [16], le BPM est un ensemble complet d'outils conçus pour comprendre, documenter, analyser et optimiser les opérations organisationnelles complexes. Son objectif principal est de superviser l'exécution des travaux afin de garantir des résultats cohérents et d'identifier les possibilités d'amélioration. La modélisation des processus, qui traduit la dynamique complexe des processus en représentations visuelles compréhensibles ou en artefacts exécutables, est au cœur de cette discipline.

2.1.1 PERSPECTIVES FONDAMENTALES DE LA MODÉLISATION DES PROCESSUS

Un modèle de processus métier holistique intègre plusieurs perspectives afin de saisir pleinement les nuances des opérations organisationnelles. La perspective du flux de contrôle est fondamentale, car elle cartographie la séquence des activités et leurs interdépendances. Sur cette structure s'appuient la perspective des données, qui régit la manière dont les données sont créées, manipulées et consommées ; la perspective des ressources, qui identifie les actifs humains et système nécessaires à l'exécution ; et la perspective opérationnelle, qui décrit les règles et la sémantique d'exécution régissant le processus [28, 30]. Alors que le BPM traditionnel met l'accent sur la perspective du flux de contrôle (cartographie des séquences d'activités), les approches modernes doivent de plus en plus intégrer la perspective des données pour gérer la complexité de la prise de décision dans le monde réel [28].

2.1.2 LE PASSAGE À UN BPM BASÉ SUR L'IA

Historiquement, la modélisation des processus a nécessité un effort manuel considérable ainsi qu'une expertise approfondie dans les langages formels, au premier rang desquels le Business Process Model and Notation [30]. Maintenu en tant que standard par l' Object Management Group (OMG), le BPMN constitue le langage de modélisation de référence dans l'industrie, offrant une notation graphique structurée pour représenter les flux d'activités, d'événements et d'unités organisationnelles au sein d'un processus métier. Son vocabulaire fondamental s'articule autour de cinq éléments clés : les activités (rectangles aux coins arrondis), les événements (cercles), les passerelles (losanges), les flux de séquence (flèches) et les pools/couloirs, délimitant les responsabilités des acteurs impliqués.

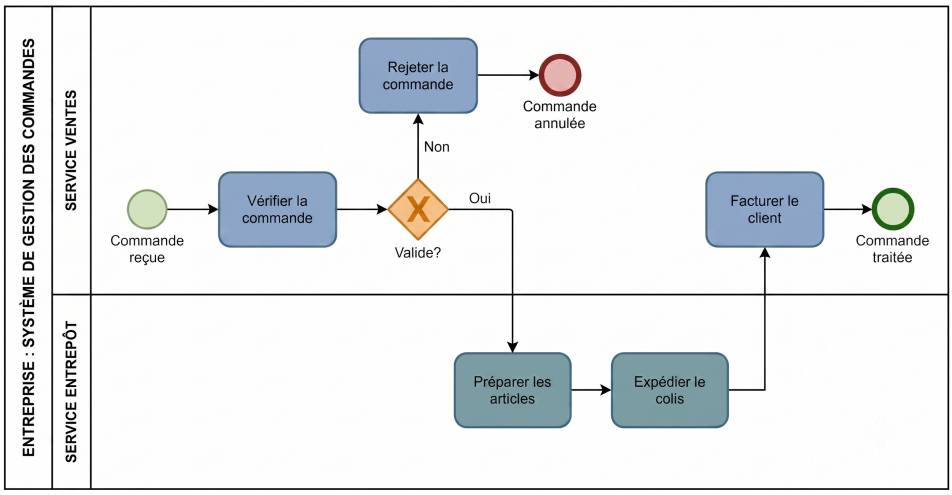

Figure 2.1 : Modèle BPMN de gestion des commandes d'une entreprise quelconque.

Afin d'illustrer concrètement la mise en œuvre de ces éléments, la figure 2.1, présente un modèle BPMN d'un système de gestion des commandes clients au sein d'une entreprise. Il intègre plusieurs services et met en évidence les interactions entre les différentes parties prenantes des ventes à l'entrepôt.

Néanmoins, la maintenance de ces modèles afin qu'ils reflètent fidèlement l'évolution des réalités opérationnelles engendre un décalage de synchronisation continu. Par ailleurs, la complexité inhérente aux processus du monde réel se traduit fréquemment par des modèles denses, difficilement appréhendables par les non-spécialistes, ce qui génère une surcharge cognitive significative [29].

Pour pallier à ces limites, le domaine connaît actuellement un changement de paradigme vers une gestion du cycle de vie basée sur l'IA. Des recherches récentes explorent l'extraction d'informations sur les processus à partir de documents textuels non structurés, tels que des

politiques ou des lignes directrices, afin d'automatiser la génération de modèles [22]. Les *LLMs* sont apparus comme des outils puissants dans ce domaine, capables de traiter des entrées textuelles complexes afin de générer et d'affiner des modèles de processus structurés. En automatisant la transformation des descriptions en langage naturel en représentations formelles des processus (par exemple *BPMN*), les *LLMs* offrent un moyen de rationaliser la modélisation des processus et de réduire la dépendance à l'égard d'un travail manuel spécialisé [28].

2.2 CORRESPONDANCE D'ENTITÉS : L'ÉVOLUTION DE L'INTÉGRATION DES DONNÉES

La correspondance d'entités ou *EM*, également appelée liaison d'enregistrements, association d'entités, résolution d'entités ou détection de doublons, est la tâche qui consiste à identifier des paires d'entrées de données qui font référence à la même entité du monde réel [27]. Il s'agit d'un défi fondamental dans l'intégration des données, essentiel pour des applications allant des soins de santé au commerce électronique, où il est crucial de déterminer si des enregistrements provenant de différentes sources représentent le même objet.

2.2.1 LE PIPELINE TRADITIONNEL : BLOCAGE ET CORRESPONDANCE

Le workflow *EM* standard comprend généralement deux phases principales, comme présenté sur la 2.2 : le blocage et la correspondance. Étant donné que la comparaison de chaque paire d'enregistrements dans deux grands ensembles de données (le produit cartésien) crée une charge de calcul quadratique, des heuristiques de blocage sont utilisées pour filtrer les non-correspondances évidentes [15]. Cela réduit ainsi l'espace de recherche à un ensemble de candidats gérable. Ensuite, la phase de correspondance utilise un modèle de classification afin de

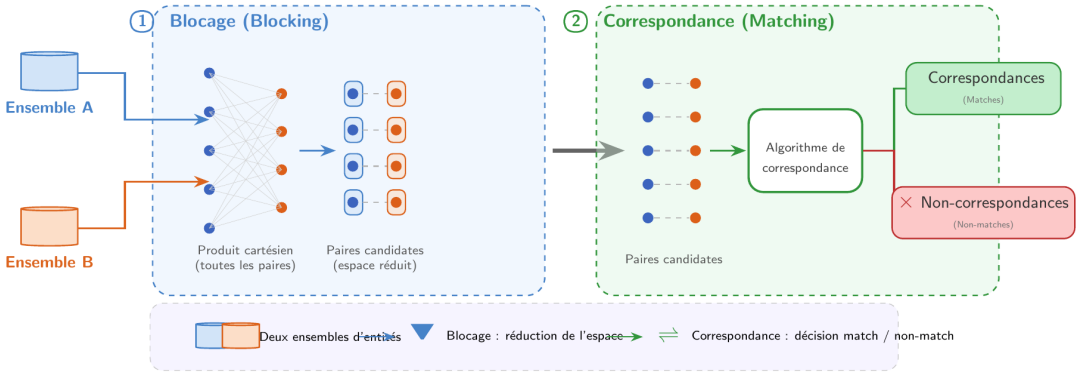

Figure 2.2 : Workflow EM standard

catégoriser les paires de candidats restantes comme correspondances ou non-correspondances. Traditionnellement, cette classification reposait sur la comparaison des similitudes textuelles entre les paires, utilisant souvent des fonctions de similarité (par exemple, Jaccard, Levenshtein) pour calculer les vecteurs de caractéristiques.

2.2.2 DES SYSTÈMES BASÉS SUR DES RÈGLES À L'APPRENTISSAGE PROFOND

Les techniques *EM* ont évolué au fil des générations. Les premières approches étaient souvent basées sur des règles, s'appuyant sur des prédicats logiques définis par des experts du domaine [41]. Elles ont été suivies par les algorithmes traditionnels de *ML* comme les modèles *Support Vector Machine (SVM)*, *RandomForests*, qui traitaient l'association d'entités comme un problème de classification binaire vecteurs de caractéristiques dérivés de mesures de similarité entre chaînes de caractères (par exemple, distance de Jaccard, distance de Levenshtein).

Les progrès récents se sont orientés vers les modèles de *Deep Learning* tels que *DeepMatcher* [36] ou *DeepER* [18]. Ces derniers utilisent des *embeddings Word2Vec* [35] ou *GloVe* [40] et

des architectures telles que les *Recurrent Neural Network (RNN)* pour capturer les similitudes sémantiques plutôt que de se fier uniquement au chevauchement syntaxique [41]. Cependant, ces méthodes *DL* nécessitent souvent de grandes quantités de données d'entraînement étiquetées et peuvent souffrir d'un manque d'interprétabilité [43]. L'état de l'art a encore progressé avec l'intégration de LLMs pré-entraînés. Des systèmes tels que *Ditto* [32] considèrent l'*EM* comme un problème de classification de paires de séquences, en s'appuyant sur des modèles basés sur l'architecture *Transformer* pour capturer des connaissances sémantiques de haut niveau et des nuances spécifiques au domaine, telles que la compréhension que *Sharp resolution* et *Sharp TV* ont des significations différentes pour le mot *Sharp*, ce que les représentations vectorielles traditionnelles négligent souvent [32].

2.2.3 CORRESPONDANCE GÉNÉRALISÉE D'ENTITÉS

Les applications concrètes nécessitent souvent une formulation plus souple, appelée *Generalized Entity Matching (GEM)*. Contrairement à la correspondance d'entités classique, qui suppose généralement des schémas homogènes (par exemple, la correspondance entre deux tables relationnelles), le *GEM* permet la liaison entre des entités de formats divers (structurés, semi-structurés ou non structurés) [46]. Par exemple, la mise en correspondance d'un CV semi-structuré avec une offre d'emploi non structurée nécessite de saisir des relations qui vont au-delà de la simple identité de l'entité, telles que la *compétence*. Cela nécessite de solides capacités de compréhension du langage pour interpréter des données hétérogènes sans attributs parfaitement alignés.

2.3 LES *LLMs* ET L'ARCHITECTURE *TRANSFORMER*

Les progrès considérables réalisés en matière de modélisation automatisée des processus et de correspondance des entités reposent sur l'architecture *Transformer*. Ce modèle d'apprentissage profond a révolutionné le Traitement Automatique du Langage Naturel (TALN) ou Natural Language Processing (NLP) en permettant une gestion efficace des dépendances à longue portée et de la parallélisation, surmontant ainsi les limites des *RNN* précédents [11].

2.3.1 L'ARCHITECTURE TRANSFORMER ET L'AUTO-ATTENTION

L'architecture *Transformer* [45], introduite par Vaswani et al., utilise un mécanisme d'auto-attention qui pondère de manière différentielle l'importance des différentes parties des données d'entrée. Ce mécanisme permet au modèle de saisir les relations contextuelles entre les *tokens*, quelque soit leur distance dans la séquence de caractères. Les réseaux de neurones de type *Transformer* se composent généralement d'un encodeur et d'un décodeur, bien que de nombreux LLMs modernes utilisent des variantes de cette structure (par exemple, uniquement un encodeur comme BERT, ou uniquement un décodeur comme GPT). La figure 2.3, tirée des travaux de Vaswani et al. [45], permet de visualiser en détail l'architecture *Transformer*.

Un paradigme essentiel rendu possible par l'architecture *Transformer* est le pré-entraînement suivi d'un *fine-tuning*. Les modèles sont pré-entraînés sur de vastes corpus de texte à l'aide d'objectifs auto-supervisés (par exemple, la prédiction du *token* suivant) afin d'apprendre des représentations linguistiques générales et des connaissances du monde. Ces modèles pré-entraînés peuvent ensuite être ajustés finement sur des ensembles de données plus petits et

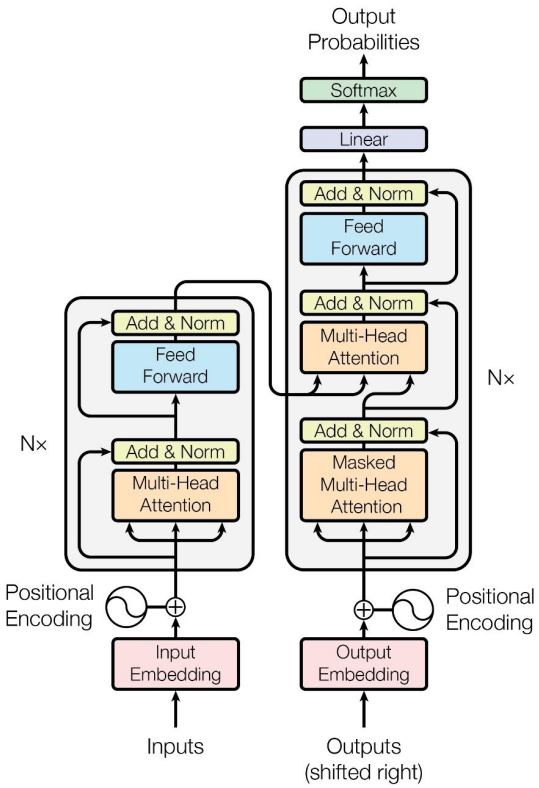

Figure 2.3 : Architecture Transformer tirée de [45]. A gauche l'Encodeur et à droite le Décodeur.

spécifiques à une tâche afin de s'adapter à des tâches précises telles que la correspondance d'entités ou la génération de code.

2.3.2 APPRENTISSAGE CONTEXTUEL ET RAISONNEMENT EN CHAÎNE DE PENSÉE

L'augmentation de la taille des modèles et des données d'entraînement a conduit à l'émergence de l'« apprentissage contextuel », qui permet aux *LLMs* d'effectuer des tâches à partir d'instructions en langage naturel ou de quelques démonstrations (*few-shot learning*) sans mettre à jour leurs paramètres [10]. Cela permet aux modèles de généraliser efficacement à de nouvelles tâches.

De plus, les *LLMs* suffisamment vastes font preuve de capacités de raisonnement. Des techniques telles que le *Chain-of-Thought prompting* poussent les modèles à générer une série d'étapes de raisonnement intermédiaires avant d'arriver à une réponse finale [49]. Cela imite l'approche de résolution des problèmes employée par l'être humain et améliore considérablement les performances sur des tâches complexes par rapport au *prompting* standard. Cependant, le Chain-of-Thought standard peut souffrir d'hallucinations et de propagation d'erreurs, car il fonctionne comme une boîte noire sans modèle interne permettant de simuler les résultats des actions ou de maintenir la cohérence de l'état [23, 52]. Pour remédier à cela, des approches récentes proposent d'intégrer des algorithmes de planification comme le *Monte Carlo Tree Search*) aux LLMs, en réutilisant le LLM à la fois comme agent de raisonnement et comme *World model* pour explorer stratégiquement les chemins de raisonnement [23].

CHAPITRE III

TRAVAUX CONNEXES

L'optimisation des processus opérationnels dans le secteur de l'énergie repose de plus en plus sur l'exploitation et la compréhension efficace des données. Pour une entreprise publique d'envergure comme Hydro-Québec, l'association des avis d'interruption avec les interruptions planifiées s'avère essentielle à la fiabilité opérationnelle et à l'exactitude des rapports de performance. Ce défi d'appariement est fondamentalement un problème d'appariement d'entités généralisé, ou Generalized Entity Matching (GEM). Contrairement à l'approche classique (Entity Matching ou EM), qui se cantonne à identifier des correspondances strictes et standardisées entre des bases de données structurées et alignées, le GEM lève ces hypothèses rigides en s'adaptant à la complexité du monde réel. Il permet d'opérer sur des formats de données hautement hétérogènes qu'ils soient structurés, semi-structurés ou totalement non structurés sous forme de texte brut, sans imposer d'alignement préalable des schémas. De surcroît, le GEM redéfinit la sémantique de l'appariement : il ne cherche plus une stricte identité entre deux entrées, mais valide si une paire d'entités satisfait une relation binaire plus souple et personnalisée (telle que la pertinence ou la qualification), formalisée par l'attribution d'une étiquette binaire à chaque paire candidate.

Dans le contexte d'Hydro-Québec, ce problème est exacerbé par la rigidité des règles opérationnelles rédigées en langage naturel, faisant de la relation d'appariement un défi sémantique complexe. Les manquements dans ce processus de correspondance compromettent directement l'intégrité du Taux de Respect des Interruptions Plannifiées (TRIP), un indicateur de performance critique censé mesurer la fiabilité du réseau et l'impact réel des travaux d'entretien sur les clients. Ce chapitre établit les fondements théoriques de notre recherche en synthétisant la littérature scientifique dans trois domaines convergents : l'intégrité des données (en explorant l'évolution de la correspondance d'entités traditionnelle vers le paradigme du GEM), le traitement automatique du langage naturel (TALN) pour décoder l'ambiguïté textuelle, et la gestion des processus métier (BPM) pour contextualiser l'application de ces méthodes aux flux opérationnels.

3.1 QUESTION DE RECHERCHE

Les systèmes traditionnels basés sur des règles fournissent une logique déterministe pour la mise en correspondance des données, mais manquent souvent de la flexibilité nécessaire pour gérer les particularités inhérentes aux situations réelles du terrain. À l'inverse, si les approches modernes d'apprentissage profond offrent une grande adaptabilité, elles fonctionnent souvent comme des « boîtes noires », sans l'interprétabilité requise pour les opérations critiques. Par conséquent, afin de relever le défi spécifique de l'optimisation du processus de mise en correspondance des notifications et des interruptions, ce mémoire pose la question de recherche ci-dessous, telle que mise en évidence par la figure 3.1 :

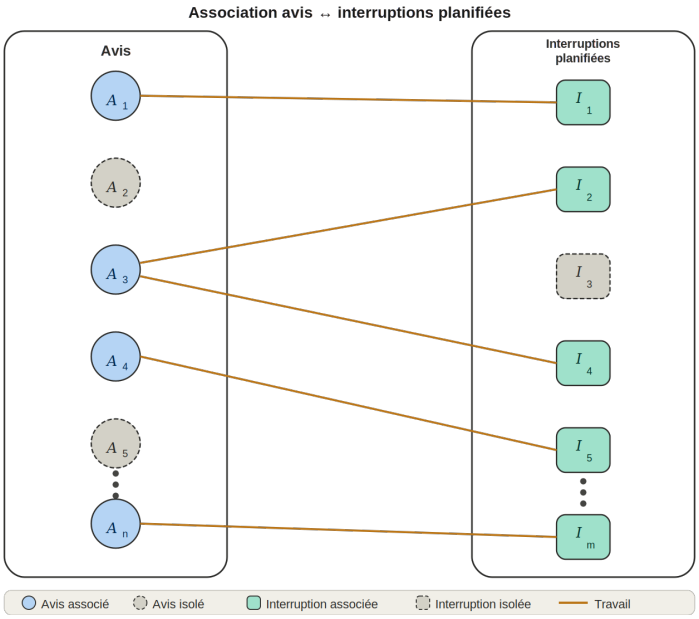

Figure 3.1 : Représentation visuelle du problème d'association des avis aux interruptions planifiées

Comment les LLMs, enrichis par les cadres d'IA explicable, peuvent-ils être exploités pour optimiser les règles d'association des avis aux interruptions planifiées dans le secteur de l'énergie, minimisant ainsi les entités non associées et améliorant la fiabilité du calcul de l'indicateur TRIP ?

3.2 STRATÉGIE DE RECHERCHE

Afin de situer cette recherche dans l'état actuel des connaissances, une revue de la littérature a été menée. La stratégie de recherche ciblait l'intersection entre la gestion des données, l'intelligence artificielle et l'optimisation des processus comme le témoigne la représentation graphique de la figure 3.2.

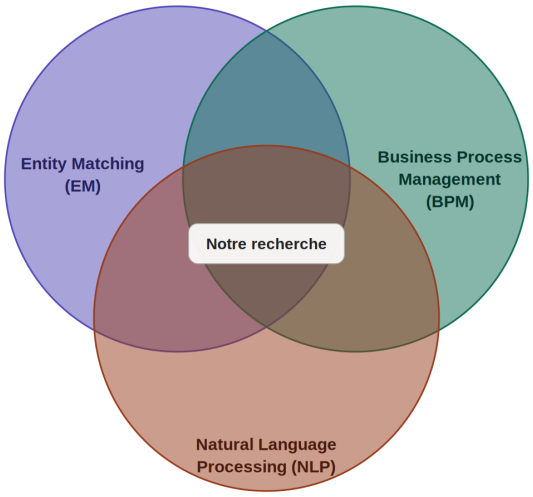

Figure 3.2 : Cadre de notre recherche

La méthodologie a utilisé les principales bases de données universitaires, notamment ACM Digital Library, IEEE Xplore et arXiv, afin de recenser à la fois les études fondamentales évaluées par des pairs et les recherches en rapide évolution dans le domaine de l'IA générative. La recherche a utilisé des combinaisons booléennes de groupes de mots-clés conçues pour combler le fossé entre la mise en œuvre technique et le contexte des processus métier :

- Intégrité et correspondance des données : « résolution d'entités », « correspondance d'entités généralisée », « liaison d'enregistrements », « résolution d'entités sans apprentissage ».
- IA et NLP : « grands modèles linguistiques », « chaîne de pensée », « raisonnement », « architectures de transformateurs ».
- Application au domaine : « Optimisation des règles métier », « Systèmes d'alimentation électrique », « Gestion des processus métier », « Génération automatisée de règles ».

- Validation et confiance : « IA explicable (XAI) », « Human-in-the-loop », « Mesures d'évaluation », « Hallucination ».

3.3 REVUE DE LA LITTÉRATURE : SYNTHÈSE THÉMATIQUE

Les sections suivantes synthétisent les travaux universitaires existants en cinq piliers thématiques pertinents pour l'optimisation du processus de mise en correspondance des avis et des interruptions.

3.3.1 L'ÉVOLUTION DE LA CORRESPONDANCE D'ENTITÉS : DES RÈGLES À LA GÉNÉRALISATION

La correspondance d'entités (EM), processus consistant à identifier les enregistrements de données qui font référence à la même entité du monde réel, est la pierre angulaire de l'intégration des données. Historiquement, les solutions EM telles que Magellan [27] se concentraient sur l'ensemble du pipeline, du blocage à la correspondance, en s'appuyant fortement sur des systèmes basés sur des règles ou des classificateurs traditionnels d'apprentissage automatique tels que les SVMs et les forêts aléatoires. Si ces systèmes offrent une meilleure interprétabilité, leur maintenance est toutefois très laborieuse et ils peinent à gérer l'hétérogénéité sémantique des textes non structurés. Des expérimentations comparatives ont d'ailleurs souligné ces limites : sur des ensembles de données textuelles et bruitées, l'approche de Magellan a plafonné à des scores F1 moyens de respectivement 83,4% et 68,5%, laissant une importante marge d'erreur [36].

Depuis, le domaine a connu un changement de paradigme vers le Deep Learning. Les premières approches neuronales, telles que DeepMatcher [32], ont démontré que les RNN pouvaient surpasser les méthodes traditionnelles sur les ensembles de données textuelles et des données contenant du bruit (*dirty data*) en capturant des dépendances séquentielles que les systèmes basés sur des règles ne parviennent pas à détecter. Les évaluations empiriques de l'architecture hybride de DeepMatcher ont par exemple affiché une amélioration moyenne remarquable de 19,4 points de pourcentage du score F1 sur des données bruitées par rapport à Magellan [36].

Plus récemment, l'état de l'art s'est cristallisé autour des LLMs pré-entraînés. Des systèmes tels que Ditto exploitent les architectures Transformer (par exemple, BERT, RoBERTa) pour présenter l'EM comme un problème de classification de paires de séquences, capturant ainsi des nuances sémantiques profondes. Cela permet notamment la distinction entre des versions spécifiques de produits ou, dans le contexte des services publics, entre différents types de pannes, que les mesures purement basées sur la distance textuelle ne parviennent souvent pas à identifier. Sur le plan technique, l'intégration par Ditto de connaissances expertes du domaine, de techniques de résumé de texte pour les attributs trop longs et de méthodes d'augmentation de données, lui a permis de surpasser l'état de l'art précédent (DeepMatcher) avec des gains atteignant jusqu'à 29% du score F1, et jusqu'à 31% sur des jeux de données particulièrement sales comme Walmart-Amazon [32]. Fait notable, les expériences ont prouvé que Ditto pouvait atteindre ces résultats optimaux en n'utilisant que la moitié, voire le cinquième, des données d'entraînement étiquetées [32].

Cependant, le défi spécifique à Hydro-Québec consiste à faire correspondre des données structurées de formats différents. Cela correspond au problème émergent du GEM (Generalized Entity Matching) que Wang et al. ont présenté Machop [46], un cadre GEM de bout en bout,

en faisant valoir que les applications du monde réel nécessitent souvent de faire correspondre des entités de formats divers (par exemple, faire correspondre des CV semi-structurés à des offres d'emploi non structurées) où la relation va au-delà de la simple identité pour s'étendre à des associations plus larges (comme la qualification d'un candidat pour un poste). Lors d'évaluations sur des ensembles de données réels d'Indeed.com, l'architecture Machop-Sequenced (qui intègre une couche de pooling sensible à la structure) a surpassé la méthode Ditto de 17,1% sur le score F1 pour la mise en correspondance d'offres d'emploi et de CV [46].

De même, *PromptEM* [47] aborde le GEM dans des environnements à faibles ressources en utilisant le *prompt-tuning*, ce qui réduit la dépendance à l'égard des grands ensembles de données étiquetées, souvent rares dans les environnements industriels. Des expériences menées dans des conditions de supervision extrêmes (par exemple en utilisant seulement 80 exemples étiquetés pour entraîner le modèle) ont démontré que *PromptEM* maintient des performances compétitives là où d'autres approches s'effondrent. Une étude d'ablation a d'ailleurs révélé que la suppression de son module de prompt-tuning entraînait une chute moyenne de 15,7% du score F1, confirmant que cette technique est indispensable pour stimuler les connaissances latentes du modèle linguistique sous-jacent [47].

Enfin, pour les scénarios où les données étiquetées sont totalement absentes, *ZeroER* [50] présente une approche de modélisation générative qui permet d'obtenir des performances compétitives en utilisant zéro exemple étiqueté. Cette solution repose sur l'analyse de la distribution des vecteurs de similarité, en partant de l'observation simple que la distribution des caractéristiques des correspondances diffère grandement de celle des non-correspondances. De manière impressionnante, ZeroER atteint des scores F1 comparables aux classificateurs

d'apprentissage supervisé (tels que la régression logistique et le perceptron multicouche) qui nécessitent pourtant que 50% des données soient annotées manuellement pour l'entraînement, économisant ainsi un effort d'étiquetage de dizaines de milliers d'exemples [50].

3.3.2 L'IA DANS LE DOMAINE DE L'Electric Power Domain

L'intégration de l'IA dans le secteur EPD transforme la gestion du réseau, même si la littérature révèle une répartition inégale des axes de recherche. Une étude exhaustive menée par Amjad et al., reposant sur l'analyse de plus de 45 études récentes, souligne que les LLMs sont principalement utilisés pour des tâches telles que la prévision de la charge, le diagnostic des pannes et les *chatbots* destinés aux clients. Les évaluations de ces implémentations mettent en évidence des bénéfices quantifiables, tels qu'une amélioration allant jusqu'à 20% de la précision des prévisions de charge et une diminution de 40% de la charge de travail manuelle pour les tâches réglementaires. Par exemple, des cadres tels que Time-LLM [25] reprogramment les données chronologiques continues en prototypes textuels discrets, ce qui permet d'exploiter les LLMs figés (sans modifier leurs poids) pour la prévision de séries temporelles. En intégrant des instructions déclaratives via une technique de Prompt-as-Prefix, les expérimentations ont démontré que Time-LLM permettait une amélioration de 20% de l'erreur quadratique moyenne ou *Root Mean Square Error (RMSE)* par rapport à des modèles traditionnels comme ARIMA [9] et LSTM [24].

Dans le domaine du soutien opérationnel, les LLMs ont été utilisés pour répondre à des questions relatives à la conformité réglementaire et aux normes de sécurité, surpassant les méthodes de recherche par mots-clés dans la récupération d'informations pertinentes. En pratique, des systèmes exploitant des modèles BERT *fine-tunés* sur des corpus de réglementations électriques

ont été développés pour traiter le sens sémantique et contextuel des requêtes. Testés sur des ensembles de données d'assurance qualité spécialisés dans l'énergie, ces modèles ont atteint une précision de réponse allant jusqu'à 92%, prouvant leur efficacité face à une terminologie de domaine très spécifique [6].

En outre, des études ont exploré l'utilisation du Natural Language Generation (NLG) pour convertir les données d'alarme brutes des éoliennes en descriptions lisibles par l'homme afin d'aider les opérateurs de maintenance. Ces systèmes utilisent des modèles basés sur l'architecture Transformer, entraînés sur des historiques d'alarmes associés à des descriptions manuelles, pour afficher en temps réel des alertes contextualisées sur les tableaux de bord. Cette traduction de données brutes en informations exploitables a contribué à une réduction allant jusqu'à 30% du temps de réponse opérationnel, minimisant ainsi les temps d'arrêt des turbines [6].

Par ailleurs, les travaux de Legris et al. de l'Université du Québec à Chicoutimi (UQAC) démontrent le potentiel critique des LLMs pour automatiser la transparence administrative chez Hydro-Québec. Face aux limites linguistiques et aux risques d'exfiltration des données inhérents aux outils commerciaux (tels que Collibra AI ou Databricks AI, souvent limités à l'anglais par défaut), les chercheurs de l'UQAC ont développé une bibliothèque Python interne à l'entreprise, sur mesure exploitant GPT-4 via afin de garantir la souveraineté des données. En structurant le contexte via le format *Database Markup Language (DBML)* pour maximiser la compréhension sémantique tout en minimisant les coûts de lexique, cette approche générative a atteint un score d'évaluation humaine probant de 90,48%, surpassant largement les systèmes génériques de l'industrie (Collibra AI plafonnant à 73,81% et Databricks AI à 71,03%) et

affichant une bien meilleure similarité cosinus (49,45%) face aux descriptions de référence [31].

Cependant, malgré des succès notables dans l'orchestration de tâches d'ingénierie et la documentation des métadonnées pour une meilleure gouvernance des données, il existe un manque de recherches sur l'optimisation des processus administratifs internes. Ce vide est particulièrement vrai pour les défis de l'intégration de données inter-systèmes, tels que le problème de la correspondance entre les avis clients et les interruptions de courant planifiées. Si les LLMs sont utilisés pour la gestion de haut niveau du réseau et les interactions avec les clients externes, leur application dans le raffinement des règles d'affaire qui régissent l'intégrité des données et le calcul des *Key Performance Indicators (KPIs)* reste sous-explorée.

3.3.3 RAISONNEMENT ET PLANIFICATION : AU-DELÀ DE LA SIMPLE GÉNÉRATION DE TEXTE

L'optimisation des règles métier nécessite plus que la simple reconnaissance de modèles ; elle exige des capacités de raisonnement pour discerner pourquoi un avis correspond à une interruption planifiée spécifique. Les progrès récents en matière de grands modèles de langage (LLMs) ont dépassé la simple génération de *tokens* pour atteindre la résolution de problèmes complexes.

Il a été démontré que les techniques de *Chain-of-Thought prompting* permettent de susciter des capacités de raisonnement considérées comme une compétence cognitive émergente liée à l'augmentation drastique de la taille des modèles (autour de 100 milliards de paramètres), leur permettant ainsi de décomposer des problèmes en plusieurs étapes en étapes intermédiaires [49]. Cette approche émule le processus de pensée humain, permet d'allouer dynamiquement

plus de calcul aux problèmes difficiles, et offre une fenêtre interprétable sur le processus décisionnel de l'IA. Afin d'améliorer la fiabilité de ces traces de raisonnement, l'auto-cohérence (*self-consistency*) agrége les résultats sur divers chemins de raisonnement afin de trouver la réponse la plus cohérente. Par exemple, sur le jeu de données de résolution mathématique *GSM8K*, l'utilisation du *Chain-of-Thought* avec le modèle *PaLM-540B* a augmenté le taux de réussite de 39 points de pourcentage (passant de 17,9% à 56,9%) par rapport au prompting standard, surpassant même des modèles (comme *GPT-3*) spécifiquement affinés pour cette tâche avec un vérificateur. Ces gains s'étendent au raisonnement de sens commun, où le *Chain-of-Thought* a permis de dépasser l'état de l'art sur *StrategyQA* (75,6% contre 69,4%) et même de battre des humains sur des tâches de compréhension sportive (95,4% de précision contre 84%) [49].

Cependant, face à des environnements interactifs, le simple raisonnement statique montre ses limites. De plus, des frameworks tels que *Reasoning + Acting (ReAct)* [52] créent une synergie entre le raisonnement et l'action , permettant aux modèles d'intercaler la génération dynamique de pensées avec des actions réelles (par exemple, interroger une base de connaissances externe pour mettre à jour leur contexte). Cette boucle de rétroaction permet de résoudre des tâches nécessitant beaucoup de connaissances empiriques, réduisant ainsi grandement les erreurs de logique et les hallucinations

Malgré ces avancées, les LLMs peinaient encore sur des tâches de planification délibérée à long terme en raison de l'absence d'une représentation interne de leur environnement. Des approches plus avancées, telles que le Reasoning viA Planning (RAP) [23], surmontent ce verrou en réutilisant les LLMs à la fois comme *World Model* et comme agents de raisonnement. S'inspirant de la cognition humaine, le *World Model* est utilisé pour anticiper et simuler l'état

futur de l'environnement ou les conséquences à long terme d'une action, sans nécessiter d'interaction avec le monde réel. L'agent de raisonnement, quant à lui, propose un éventail d'actions possibles face à cet état donné. En utilisant la recherche arborescente Monte Carlo couplée à des fonctions de récompense (par exemple l'évaluation de la probabilité d'une action ou de la confiance en un état), le framework RAP explore stratégiquement les chemins de raisonnement sous la forme d'un arbre. Cette planification permet un équilibre rigoureux entre l'exploration de nouvelles pistes de réflexion non visitées et l'exploitation des étapes de raisonnement ayant généré les meilleures récompenses historiques.

Ces méthodologies suggèrent que les LLMs peuvent être transformés de générateurs de texte passifs en agents actifs capables d'exécuter et de valider une logique métier complexe.

3.3.4 L'IA DANS LE BPM

Le BPM évolue de la modélisation manuelle vers l'automatisation basée sur l'IA. Les recherches menées par Kourani et al. démontrent que les LLM peuvent automatiser la génération de modèles de processus (par exemple, BPMN, réseaux de Petri) directement à partir de descriptions en langage naturel. Cette capacité permet de traduire des protocoles opérationnels non structurés en une logique formelle exécutable. De plus, des cadres tels que DENAS [12] ont été proposés pour combler le fossé entre les réseaux de neurone et les systèmes basés sur des règles en automatisant la génération de règles via l'extraction de connaissances à partir de réseaux de neuronaux profonds.

Cependant, l'application des LLMs dans le BPM n'est pas sans limites. Si les LLMs peuvent effectuer des tâches telles que l'exploration de modèles déclaratifs ou l'évaluation de l'adéquation au RPA, ils nécessitent une comparaison minutieuse avec les outils spécialisés existants.

La littérature suggère une évolution vers l'utilisation des LLMs non seulement pour la modélisation, mais aussi pour affiner activement la logique au sein des processus établis, bien que ce domaine soit moins développé que la simple génération de modèles [22].

3.3.5 EXPLICABILITÉ, ÉQUITÉ ET ÉVALUATION

Dans les secteurs réglementés tels que celui de l'énergie, les décisions de type « boîte noire » sont inacceptables ; la confiance et la responsabilité sont primordiales. L'eXplainable Artificial Intelligence est donc une exigence essentielle. Pour la correspondance d'entités en particulier, les méthodes XAI standard telles que LIME [42] échouent souvent en raison des effets d'interaction entre les enregistrements. LEMON [8] résout ce problème en produisant des explications doubles et en introduisant le « potentiel d'attribution » pour expliquer efficacement les non-correspondances. De même, Mojito [14] adapte LIME spécifiquement aux modèles d'association basés sur le Deep Learning afin de produire des explications interprétables. D'autres outils comme EXPLAINER [17] fournit un cadre permettant de comprendre et d'analyser les modèles associatifs, ce qui permet aux utilisateurs de comprendre pourquoi certaines paires d'entités sont appariées ou non.

Au-delà de l'explicabilité, l'équité dans le matching est une préoccupation croissante. Shahbazi et al. soulignent que les techniques de EM peuvent présenter un biais à l'égard des groupes sous-représentés, ce qui nécessite un contrôle rigoureux de l'équité.

Enfin, la validation des résultats des modèles génératifs nécessite des mesures robustes. Les mesures traditionnelles n-gram telles que *BLEU* [39], *ROUGE* [33] sont insuffisantes pour évaluer le raisonnement ou l'équivalence sémantique. Sur le plan technique, ces métriques reposent strictement sur le chevauchement des mots entre le texte prédit et le texte de référence.

Par conséquent, elles échouent à reconnaître la diversité compositionnelle et lexicale qui préserve le sens, pénalisant injustement les paraphrases valides. Plus grave encore pour les processus métier, les modèles basés sur les *n-gram* peinent à capturer les dépendances lointaines : par exemple, un simple changement d'ordre des mots inversant une cause et son effet ne sera que très faiblement pénalisé par *BLEU* si la fenêtre d'analyse (la taille du n-gram) est petite. De même, *ROUGE*, bien qu'orienté vers le rappel, ne fournit aucune information sur le flux narratif, la grammaire, la structure logique ou l'exactitude factuelle du texte généré. Ces limites inhérentes expliquent leur faible corrélation avec le jugement humain lorsque la qualité sémantique des modèles augmente.

Pour surmonter ces contraintes, de nouvelles mesures telles que BERTScore [53] s'affranchissent de la correspondance exacte de chaînes de caractères. BERTScore calcule une similarité continue en effectuant la somme des des similarités cosinus entre les représentations vectorielles contextuelles (contextual embeddings) des tokens générés et de référence . Cette approche permet de capturer efficacement les dépendances à longue distance et les nuances sémantiques profondes.

Cependant, pour les tâches de raisonnement multi-étapes (comme le CoT), évaluer le texte dans son ensemble reste insuffisant, car les métriques générales de génération ne sont pas conçues pour mesurer le gain d'information ou les incohérences logiques isolées au sein des étapes de réflexion. C'est pourquoi *ROSCOE* [21] propose une suite de mesures permettant d'évaluer les raisonnements étape par étape, en vérifiant la cohérence sémantique et les erreurs d'inférence logique. Techniquement, ROSCOE décompose son évaluation selon quatre perspectives : (1) l'alignement sémantique, qui vérifie vectoriellement si chaque étape de raisonnement est correctement ancrée dans le contexte source ; (2) l'inférence logique, qui utilise des modèles

de NLG pour détecter les probabilités de contradictions internes et d'erreurs de déduction ; (3) la similarité sémantique, conçue pour identifier les hallucinations ou les répétitions inutiles ; et (4) la cohérence linguistique globale Ces mesures avancées sont essentielles pour valider les « processus de réflexion » d'un optimiseur de règles métier basé sur de l'IA.

3.3.6 RÉSUMÉ

En conclusion, cette revue de la littérature met en évidence une convergence technologique fondamentale nécessaire pour répondre aux défis d'intégration et d'optimisation au sein des organisations modernes. D'une part, l'évolution de la correspondance d'entités (Entity Matching) montre une transition claire des méthodes historiques basées sur des règles vers des approches hybrides intégrant apprentissage automatique, apprentissage profond et correspondance d'entités généralisée (Generalized Entity Matching). Ces avancées permettent désormais de réconcilier des formats de données hétérogènes (structurés et non structurés) tout en capturant des relations complexes et contextuelles, même dans des environnements où les données étiquetées sont limitées. D'autre part, l'intégration de l'intelligence artificielle dans la gestion des processus métier (Business Process Management) transforme des processus statiques en systèmes dynamiques, adaptatifs et continuellement optimisés, capables d'intégrer des mécanismes de décision intelligents au cœur des opérations.

Dans le secteur de l'énergie électrique (EPD), ces avancées ouvrent des perspectives particulièrement pertinentes pour des problématiques encore peu explorées, telles que l'association entre les avis d'interruptions destinés aux clients et les interruptions planifiées effectivement réalisées sur le réseau. Ce problème peut être interprété comme un cas spécifique de correspondance d'entités dans un contexte industriel contraint, caractérisé par des données fragmentées, des

relations implicites et des règles métier complexes. Contrairement aux tâches classiques d'EM, cette association repose sur une combinaison de critères temporels, géographiques, opérationnels et contextuels. La littérature actuelle n'aborde que marginalement ce type de problématique, laissant un espace de recherche pour des approches hybrides capables de combiner robustesse des règles existantes et capacité d'apprentissage des modèles.

Dans ce contexte, les LLMs offrent une opportunité nouvelle, non pas uniquement comme outils de génération de texte, mais comme mécanismes de formalisation et d'enrichissement des règles métier. L'intégration de techniques de raisonnement telles que le Chain-of-Thought, les cadres action-réflexion (ReAct) ou encore les approches de planification (RAP) permet d'envisager ces modèles comme des agents capables d'analyser des cas complexes, d'identifier des régularités et de proposer des extensions cohérentes aux procédures existantes. Appliqués au problème d'association des avis et des interruptions, ces modèles peuvent ainsi servir d'interface entre les connaissances issues des données (via le Machine Learning et l'XAI) et leur traduction en règles opérationnelles interprétables.

Cependant, l'introduction de telles approches dans un domaine critique comme celui de la distribution d'électricité nécessite des garanties fortes en matière de fiabilité, d'explicabilité et de validation. La littérature souligne les limites des métriques classiques pour évaluer des systèmes fondés sur le raisonnement, et met en avant l'importance d'approches combinant évaluation quantitative (performance de classification, analyse des erreurs) et qualitative (validation humaine des règles générées). Dans ce cadre, les méthodes d'IA explicable (XAI), ainsi que l'analyse de cas spécifiques tels que les faux positifs ou les situations ambiguës, deviennent essentielles pour comprendre, justifier et affiner les mécanismes d'association proposés.

Ainsi, ce mémoire s'inscrit à l'intersection de ces différents axes en proposant d'explorer une approche hybride pour améliorer l'association entre avis d'interruptions et interruptions planifiées. Il vise à tirer parti des complémentarités entre règles métier existantes, modèles de Machine Learning, techniques d'IA explicable et capacités génératives des LLMs, afin d'identifier de nouveaux critères d'association, d'enrichir les procédures actuelles et de mieux comprendre les mécanismes sous-jacents à ces correspondances. Cette démarche, à visée exploratoire, contribue à combler un manque dans la littérature en proposant une application concrète et intégrée de ces technologies à un problème industriel réel, tout en posant les bases pour de futurs systèmes d'aide à la décision plus robustes, explicables et adaptatifs.

CHAPITRE IV

MÉTHODOLOGIE DE LA RECHERCHE

Ce chapitre présente le cadre méthodologique retenu pour répondre à la question centrale de cette recherche mentionnée plus haut qui était de savoir comment les LLMs, enrichis par les cadres d'IA explicable (XAI), peuvent-ils être exploités pour optimiser les règles d'association des avis aux interruptions planifiées dans le secteur de l'énergie, minimisant ainsi les entités non associées et améliorant la fiabilité du calcul de l'indicateur Taux de Respect des Interruptions Plannifiées (TRIP).

4.1 INTRODUCTION

En raison du caractère exploratoire de ce travail, la démarche adoptée ne vise pas à confirmer des hypothèses prédéfinies, mais à investiguer de nouvelles pistes et à mettre en évidence des relations latentes que les procédures existantes n'auraient pas encore formalisées. Cette posture, proche d'une logique de découverte, justifie le recours à une méthodologie plurielle, combinant des techniques issues du traitement des données, de l'apprentissage automatique, de l'intelligence artificielle explicable et des modèles de langage.

La démarche se déploie en cinq étapes complémentaires et séquentielles dont la vue globale est résumée à la Figure 4.1. La première étape est consacrée à la constitution et à l'enrichissement du jeu de données à partir des associations déjà réalisées par la procédure existante, auxquelles sont adjoints des cas négatifs générés par permutation contrôlée, afin d'obtenir un corpus équilibré et représentatif. La deuxième étape concerne l'entraînement et la sélection de modèles de Machine Learning supervisés, dont les performances sont évaluées principalement au moyen du F1-score. La troisième étape mobilise des méthodes d'intelligence artificielle explicable, notamment SHAP (SHapley Additive exPlanations), afin d'interpréter les décisions du modèle retenu et d'identifier, parmi les faux positifs, des signaux potentiellement révélateurs de critères d'association non encore intégrés dans la logique actuelle. La quatrième étape exploite ces explications comme matière première pour la génération, via plusieurs grands modèles de langage, de nouvelles versions enrichies de la procédure d'association. Enfin, la cinquième étape propose une évaluation à deux niveaux : une évaluation technique quantitative à travers les métriques automatiques, et une évaluation qualitative par des experts métiers de la conduite réseau, permettant de juger à la fois la pertinence opérationnelle et la clarté des propositions générées.

4.2 CONSTRUCTION DU JEU DE DONNÉES

La construction d'un jeu de données de qualité constitue une étape fondatrice de toute démarche d'apprentissage automatique supervisé, conditionnant directement la capacité des modèles à généraliser de manière fiable. Dans le cadre de cette recherche, cette étape revêt une importance particulière dans la mesure où aucun jeu de données préexistant ne couvre la problématique de correspondance d'entités dans le domaine de la gestion des réseaux de distribution électrique. Il

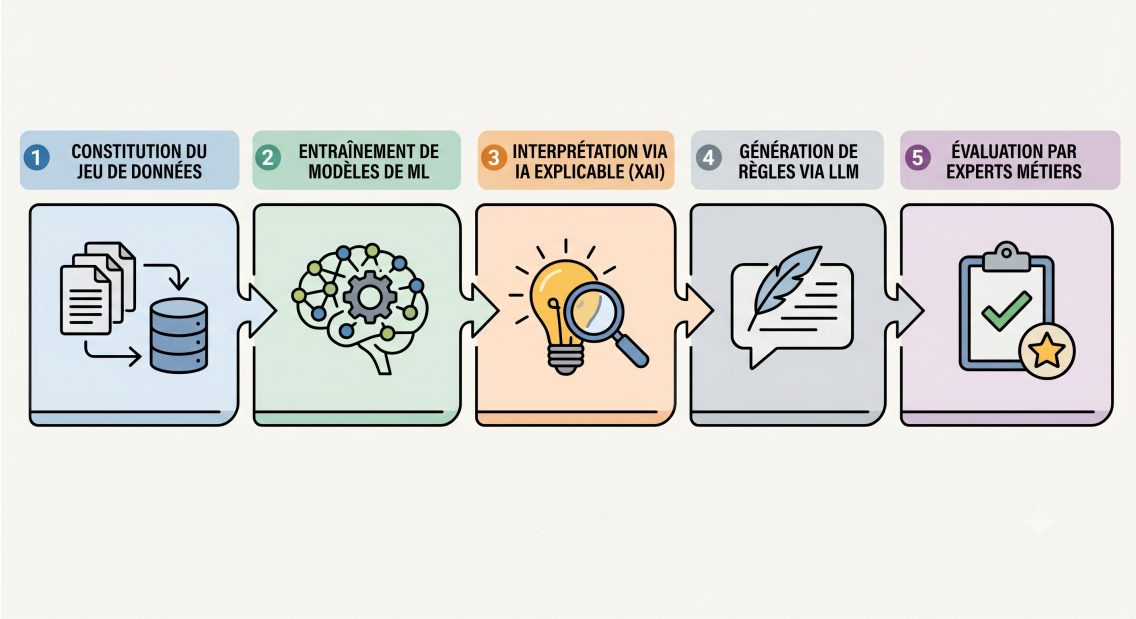

Figure 4.1 : Pipeline méthodologique de la recherche pour l'optimisation des associations avis - interruptions planifiées

a donc été nécessaire de concevoir intégralement un jeu de données ad-hoc, à partir de données opérationnelles réelles, en suivant une démarche structurée articulée autour de quatre phases successives : l'identification et l'extraction des données sources, l'enrichissement sémantique des entités, la génération des exemples négatifs et l'équilibrage du jeu de données final. Les sous-sections suivantes décrivent en détail chacune de ces étapes.

4.2.1 CONTEXTE ET SOURCE DES DONNÉES

Cette recherche a été conduite au sein d'Hydro-Québec, société d'État québécoise assurant la production, le transport et la distribution d'électricité à l'échelle de la province. L'entreprise s'appuie sur six (6) Centre d'Exploitation et de Distribution (CED) que nous désignerons par A, B, C, D, E et F ; chacun correspondant à une délimitation géographique distincte du territoire québécois. Ces centres coordonnent les équipes d'intervention sur le terrain et assurent la

gestion des rétablissements lors des pannes. Les opérateurs y utilisent un système source désigné sous l'appellation logiciel CED, permettant la manipulation de données hétérogènes relatives aux avis de coupure, aux pannes, aux interruptions planifiées, aux clients, aux conditions météorologiques, aux transformateurs, entre autres. Ces données sont ensuite agrégées, après traitement, dans un entrepôt de données centralisé, consolidant l'ensemble des informations provenant des différents CEDs.

Deux entités métier constituent le cœur de cette étude : les avis d'interruption, qui représentent des notifications ou prévisions d'arrêt du réseau, et les interruptions planifiées, qui correspondent aux actions de coupure effectivement réalisées sur ce réseau. L'association correcte entre ces deux entités est la condition nécessaire au calcul du Taux de Respect des Interruptions Plannifiées (TRIP), indicateur suivi par la direction générale et ayant un impact direct sur le bonus des techniciens, allocation de ressources, et les reporting réglementaires.

Les données exploitées sont extraites de plusieurs tables de cet entrepôt, à partir notamment de la table source de l'environnement de production contenant les associations déjà réalisées entre avis et interruptions. Des jointures sont ensuite opérées sur les tables auxiliaires relatives aux avis, aux interruptions, aux clients et aux équipements, afin d'enrichir la représentation de chaque entité.

4.2.2 STRUCTURE INITIALE ET ENRICHISSEMENT DU JEU DE DONNÉES

Dans son état initial, la table des associations comprend plus de 100000 lignes et trois (3) colonnes, couvrant une période allant de 2022 à 2025. Ces colonnes sont les suivantes : *SK_Travail* (identifiant unique de chaque travail), *SK_AVIS* (identifiant de l'avis associé) et *SK_Interruption* (identifiant de l'interruption planifiée associée). Le travail ici étant, une

Tableau 4.1 : Structure de la table source contenant les associations

SK_Travail	SK_AVIS	SK_Interruption
1	0	98638
2	1	*null*
...	...	...
126	154	*null*
...	...	...
9053	13929	*null*
...	...	...
70673	-1	146330
...	...	...
71169	-1	192079
...	...	...
83324	146716	32095
...	...	...

association entre un avis et une ou plusieurs interruptions, un avis sans interruption ou une interruption réalisée sans avis, il convient de noter que l'attribut SK_AVIS peut prendre les valeurs *-1* ou *-2*, indiquant l'absence d'avis associé à un travail, tandis que l'attribut *SK_Interruption* peut être nul (valeur *null*), signifiant l'absence d'interruption planifiée associée. Le Tableau 4.1 ci-contre présente un apercu de la table source en question.

Ces colonnes étant de nature purement identificatrice, elles ne permettaient pas d'extraire une information discriminante suffisante pour la tâche de correspondance d'entités (Entity Matching ou EM). Afin de pallier cette limitation, un processus d'enrichissement des données a été entrepris. Un document structuré recensant une liste de colonnes candidates, organisées selon plusieurs axes (caractéristiques des clients, des équipements, etc.), a été soumis à des experts du domaine aux fins de priorisation. Cette démarche visait à garantir que les attributs retenus présentent une pertinence métier avérée.

4.2.3 GÉNÉRATION DES EXEMPLES NÉGATIFS

La table source contenant d'une part les entités individuelles et d'autre part les associations avérées constituant ainsi des exemples positifs (*matchs*) résultant de l'application stricte des règles métier, il a été nécessaire de constituer des exemples négatifs (cas de non-association ou *no-matchs*) afin de formuler la tâche de correspondance d'entités comme un problème de classification binaire, conformément à la définition établie dans la littérature.

La génération de ces exemples négatifs a été réalisée en exploitant les entités non-appariées de la table source, c'est-à-dire les avis dont l'identifiant est inférieur à zéro et les interruptions dont l'identifiant est nul. En appliquant une logique de permutation sur cet ensemble d'entités libres, tout en omettant délibérément l'un des critères d'association, des paires de non-correspondance ont été synthétisées.

Afin de permettre l'apprentissage supervisé, une étape de labellisation a été réalisée par l'ajout d'une colonne binaire supplémentaire désignée *is_match*. Cette variable cible encode la classe de chaque paire : la valeur *True* est attribuée aux associations positives, c'est-à-dire aux paires dont la correspondance entre un avis et une interruption est avérée, tandis que la valeur *False* est assignée aux associations négatives, représentant les paires synthétiquement obtenues pour lesquelles aucune correspondance réelle n'existe. Cela s'illustre dans le Tableau 4.2 où il est important de noter les permutations, par rapport au Tableau 4.1, ayant engendré la création des cas non-matchs. Cette colonne constitue ainsi la variable de sortie du modèle de classification binaire entraîné dans les étapes ultérieures.

Tableau 4.2 : Cas positifs et négatifs

SK_AVIS	SK_Interruption	is_match
0	98638	True
...	...	...
154	192079	False
...	...	...
13929	146330	False
...	...	...
146716	32095	True
...	...	...

4.2.4 ÉQUILIBRAGE ET FINALISATION DU JEU DE DONNÉES

Une fois les paires d'identifiants constituées pour les cas positifs (*matchs*) et négatifs (*no-matchs*), les jointures sur les tables auxiliaires ont été appliquées afin d'enrichir chaque paire avec les attributs préalablement sélectionnés. Les observations comportant des valeurs manquantes ont été supprimées de façon à limiter le jeu de données et d'en assurer sa complétude.

Par ailleurs, afin d'obtenir un jeu de données équilibré, condition nécessaire pour éviter un biais d'apprentissage favorisant la classe majoritaire, la distribution des *no-matchs* a été calquée sur la répartition proportionnelle des *matchs* par CED. Par exemple, si le CED A représente 45% des associations positives, une proportion équivalente de 45% d'exemples négatifs lui est attribuée, et ce de manière aléatoire, pour chacun des CEDs considérés.

À l'issue de ce processus, la table initiale de 3 colonnes et plus de 100000 lignes aboutit à un jeu de données final composé d'environ 8500 lignes et 30 colonnes, incluant la variable cible is_match. Pour des raisons de confidentialité, l'ensemble des attributs retenus accompagnés de

leurs types respectifs et d'une valeur illustrative extraite du jeu de données ont été expressement omis.

4.3 ENTRAÎNEMENT DES MODÈLES

La tâche d'association entre avis d'interruption et interruptions planifiées étant fondamentalement appréhendée comme un problème d'Entity Matching (EM), elle a été formalisée comme une tâche de classification binaire résolue par apprentissage supervisé. Le jeu de données final précédemment constitué a été partitionné selon une répartition classique de 80% pour l'ensemble d'entraînement et 20% pour l'ensemble de test, conformément aux pratiques établies dans la littérature.

L'entraînement des modèles a été réalisé via AutoML. AutoML désigne une approche d'optimisation automatisée du processus d'apprentissage automatique (Machine Learning ou ML), permettant l'entraînement simultané de plusieurs modèles ML sur un même jeu de données tout en procédant à une recherche automatique des hyperparamètres optimaux pour chacun d'eux. Cette approche présente l'avantage de réduire le biais de sélection lié à une configuration manuelle des modèles, tout en permettant une exploration systématique et reproductible de l'espace des solutions.

Dans ce cadre, cinq (5) algorithmes de classification ont été évalués : les arbres de décision (*Decision Trees*), les forêts aléatoires (*Random Forests*), la régression logistique (*Logistic Regression*) via la librairie *scikit-learn*, ainsi que les algorithmes de gradient boosting *LightGBM* et *XGBoost*. Le critère de sélection du modèle optimal repose sur le F1-score, métrique harmonisant précision et rappel.

Tableau 4.3 : Performance des modèles

Modèle	F1-score (%)	Nombre de Faux positifs
XGBoost	92.15	88
LightGBM	91.13	78
Logistic Regression	83.33	137
Random Forest	74.22	353
Decision Tree	70.23	322

À l'issue de la phase d'entraînement, une attention particulière a été portée aux faux positifs produits par le meilleur modèle retenu, c'est-à-dire les paires incorrectement classifiées comme des associations valides. Ces cas constituent en effet une source d'information précieuse pour les étapes ultérieures du pipeline méthodologique. Un total de 88 instances de faux positifs a ainsi été extrait et sauvegardé dans une table dédiée, en vue de leur exploitation dans la phase suivante. Une vue d'ensemble des performances obtenues par l'ensemble des modèles est présentée dans le Tableau 4.3.

4.4 INTERPRÉTABILITÉ DES MODÈLES (XAI)

À l'issue de l'évaluation du modèle sur le jeu de test, l'analyse des résultats révèle la présence de 88 instances classifiées comme faux positifs. Ces cas, bien qu'erronément associés par le modèle, revêtent dans le cadre de cette recherche une valeur analytique particulière. En effet, notre hypothèse centrale repose sur l'idée que les règles d'association actuellement définies dans le domaine ne couvrent pas exhaustivement l'ensemble des configurations d'entités susceptibles d'être liées. Dès lors, les instances pour lesquelles le modèle prédit une association positive, en dépit de l'absence de règle explicite les validant, constituent

des candidats potentiels à l'émergence de nouvelles règles latentes. Ces associations, jugées négatives selon les critères conventionnels mais reconnues comme positives par le modèle de classification, pourraient en effet receler des critères discriminants non encore formalisés, que le modèle a appris de manière implicite à partir des données d'entraînement.

Afin d'exploiter rigoureusement ces faux positifs et d'élucider les mécanismes décisionnels sous-jacents, il est nécessaire de recourir aux outils de l'Intelligence Artificielle Explicable (*eXplainable Artificial Intelligence* ou XAI). Cette discipline se définit comme l'ensemble des systèmes et méthodologies permettant de rendre intelligibles le fonctionnement interne d'un modèle d'apprentissage automatique ainsi que ses frontières de décision, pour un auditoire donné. Au-delà de la simple transparence algorithmique, la XAI offre une dimension essentielle : les explications générées permettent aux experts du domaine de confronter la logique apprise par le modèle aux connaissances établies, afin d'en déduire des liens de causalité potentiels et d'évaluer la capacité de généralisation du modèle à des données non observées.

Parmi les méthodes XAI disponibles, notre choix s'est porté sur SHAP (SHapley Additive exPlanations), une méthode post-hoc et *model-agnostic* qui quantifie la contribution individuelle de chaque variable d'entrée à la prédiction du modèle, en s'appuyant sur le concept de valeur de Shapley issu de la théorie des jeux coopératifs. Ce choix est motivé par plusieurs avantages théoriques et pratiques déterminants par rapport à des alternatives comme LIME [42].

Sur le plan des garanties théoriques, SHAP constitue un cadre unifié pour les méthodes d'attribution additive de features, et il est démontré mathématiquement qu'il est le seul modèle d'explication de cette classe à satisfaire simultanément trois propriétés fondamentales : (i) la précision locale, garantissant que l'explication reproduit fidèlement la sortie du modèle original pour toute instance considérée ; (ii) l'absence de contribution nulle, stipulant que les

variables absentes de l'entrée originale ne se voient attribuer aucune contribution ; (iii) la cohérence, assurant que si la contribution marginale d'une variable augmente à la suite d'une modification du modèle, son attribution ne peut décroître [34]. Ces propriétés confèrent à SHAP une robustesse formelle absente dans d'autres approches. LIME, notamment, repose sur des paramètres heuristiques tels que les noyaux de pondération ou les fonctions de perte dont le choix arbitraire entraîne des violations des propriétés de précision locale et de cohérence, pouvant conduire à des explications trompeuses ou contre-intuitives. Par ailleurs, des études empiriques ont montré que les importances de variables attribuées par SHAP présentent une concordance significativement plus élevée avec le raisonnement humain comparativement à LIME. Enfin, bien que le calcul exact des valeurs de Shapley soit computationnellement prohibitif, SHAP introduit des approximations efficaces notamment via Kernel SHAP, qui exploite une régression linéaire pondérée spécialisée permettant d'obtenir des estimations précises avec un nombre d'évaluations du modèle considérablement réduit par rapport aux méthodes d'échantillonnage classiques.

Afin de visualiser cette hiérarchie globale des variables et la direction de leur impact sur les prédictions, la Figure 4.2 présente le diagramme obtenu par SHAP sur nos données. Cette représentation offre une synthèse explicite de la distribution des valeurs de Shapley : elle permet d'observer, pour quelques variables anonymisées volontairement, la corrélation entre la valeur propre de la caractéristique (élevée ou faible) et son influence, positive ou négative, sur la classification finale du modèle.

L'application de SHAP à l'ensemble des instances du jeu de données produit, pour chacune d'elles, un vecteur de valeurs de Shapley quantifiant la pondération de chaque variable dans la décision de classification. Ces vecteurs constitueront la matière première de l'étape suivante

du pipeline méthodologique : ils seront structurés sous forme de prompts destinés à un modèle de langage, en vue de la génération automatique de nouvelles règles d'association candidates.

4.5 GÉNÉRATION DE NOUVELLES RÈGLES PAR LLM

La génération de nouvelles règles d'association constitue l'étape culminante du pipeline méthodologique développé dans ce travail. Après avoir extrait des associations statistiquement significatives et identifié les variables explicatives déterminantes à l'aide des méthodes décrites dans les sections précédentes, il s'agit désormais de capitaliser sur ces résultats pour produire des règles inédites, formulées en langage naturel et ancrées dans la réalité opérationnelle du réseau de distribution d'Hydro-Québec. Pour ce faire, nous faisons appel aux grands modèles de langage ou LLMs, et plus précisément à leur capacité de raisonnement structuré, afin de synthétiser l'ensemble des connaissances accumulées en règles exploitables. Cette section décrit successivement le choix et le déploiement des modèles retenus, puis la démarche d'ingénierie des prompts mise en œuvre pour orienter efficacement leur raisonnement.

4.5.1 RECOURS AUX MODÈLES DE RAISONNEMENT

Cette étape mobilise une catégorie spécialisée de LLM connue sous le nom de modèles de raisonnement (*reasoning models*). Un modèle de raisonnement constitue une forme évoluée de LLM, conçu pour accomplir des tâches cognitives complexes en reproduisant des processus de réflexion analogues à ceux de l'être humain. Contrairement aux LLMs standards, qui s'appuient sur une génération autorégressive de tokens produisant une réponse immédiate, les modèles de raisonnement introduisent le concept de pensée intermédiaire (*thought* en anglais) :

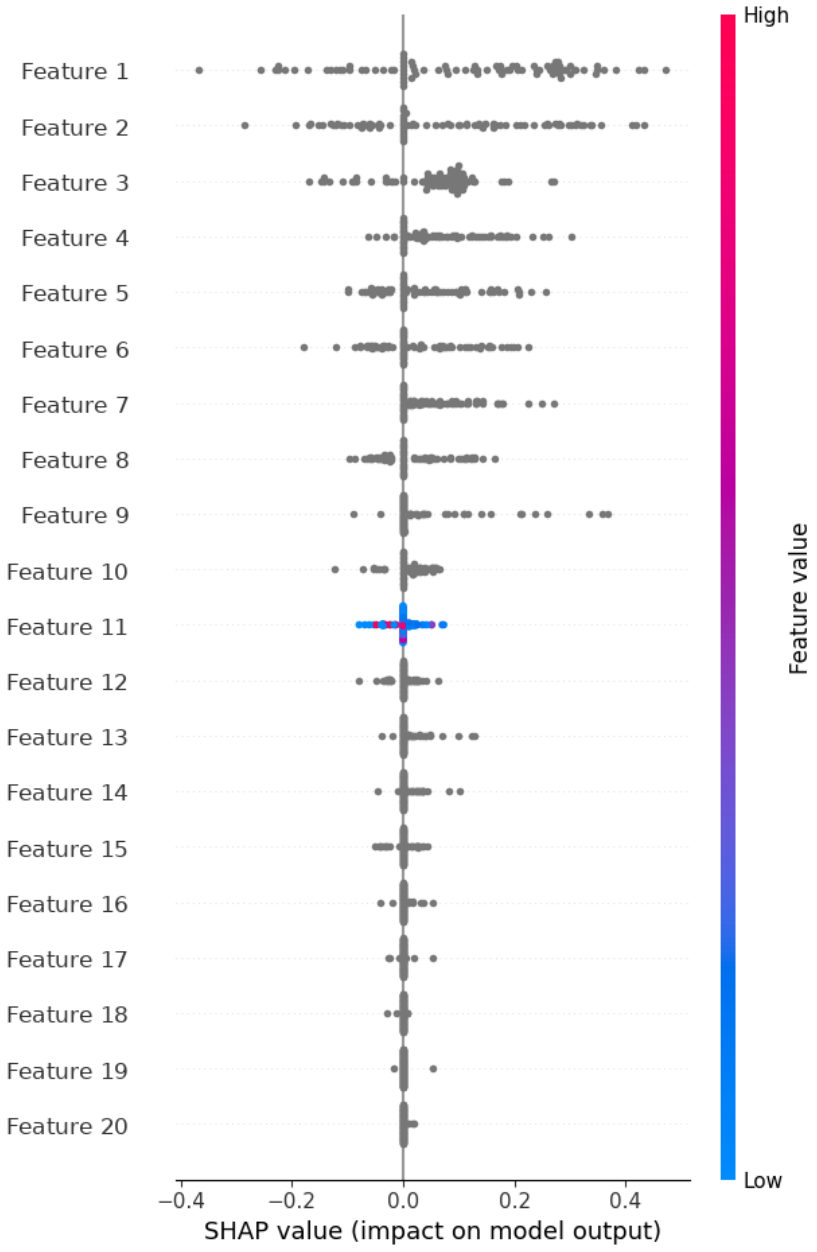

Figure 4.2 : Visualisation SHAP de l'impact des variables sur la classification - noms des variables anonymisés

une séquence de tokens représentant les étapes successives d'un processus de raisonnement. Ces modèles sont également caractérisés par leur utilisation de techniques telles que le Chain-of-Thought (CoT), qui consiste à générer une série cohérente d'étapes de raisonnement en langage naturel conduisant à une réponse finale. Plutôt que de produire immédiatement une prédiction, le modèle décompose le problème à la manière d'un expert humain, procédant étape par étape pour résoudre des problèmes à plusieurs niveaux d'abstraction.

Les principaux avantages des modèles de raisonnement qui justifient leur adoption dans ce travail sont les suivants :

- Performance supérieure sur les tâches complexes : les modèles de raisonnement démontrent des améliorations significatives dans des domaines où les LLMs traditionnels peinent, atteignant un niveau expert en mathématiques avancées, en programmation algorithmique et en raisonnement scientifique multidisciplinaire.
- Décomposition systématique des problèmes : ces modèles excellent à subdiviser des problèmes complexes en sous-tâches gérables, à identifier les principes-clés, à construire des raisonnements formels et à intégrer des connaissances issues de domaines multiples.
- Interprétabilité et auditabilité accrues : en générant explicitement une chaîne de pensée intermédiaire, les modèles de raisonnement offrent une fenêtre transparente sur leur processus décisionnel, ce qui permet de vérifier la logique employée, d'auditer les décisions produites et d'identifier d'éventuelles failles dans le raisonnement.

Ces propriétés rendent les modèles de raisonnement particulièrement adaptés à la génération de règles d'association dans un contexte métier aussi exigeant que la distribution électrique, où la rigueur et la traçabilité du raisonnement sont essentielles.

4.5.2 SÉLECTION ET DÉPLOIEMENT DES MODÈLES

Pour les besoins de cette recherche, trois modèles développés par OpenAI ont été explorés : *o3-mini*, *o1* et *o4-mini*. Le choix de ces modèles repose sur plusieurs critères complémentaires. D'une part, ils sont explicitement présentés par OpenAI comme des modèles de raisonnement, et leurs performances sur les benchmarks de référence les situent parmi les meilleurs de leur catégorie à coté de tant d'autres, ainsi que le résume le Tableau 4.4 [44, 1, 2, 3, 38]. D'autre part, ces modèles présentent un avantage économique notable par rapport à d'autres solutions offrant des performances comparables et sortis à la même période, comme l'illustre le Tableau 4.5.

Tableau 4.4 : Comparaison des performances des modèles de raisonnement à travers des benchmarks clés

	o1	**o3-mini**	**o4-mini**	**GPT-5**	**Gemini 2.5 Pro**	**Claude Opus 4**	**Claude Sonnet 4**
GPQA-Diamond	78.0%	77.0%	81.4%	–	**83.0%**	79.6%	75.4%
MMMLU	–	–	–	–	–	**88.8%**	86.5%
SWE-bench	48.9%	49.3%	68.1%	**74.0%**	63.2%	72.5%	72.7%
AIME 2025	79.2%	86.5%	92.7%	**99.6%**	83.0%	75.5%	70.5%

Tableau 4.5 : Tarification des modèles de raisonnement par million de tokens (en dollars)

LLM	**Fournisseur**	**Date de sortie**	**Input**	**Output**
o4-mini	OpenAI	Avril 2025	**1.10**	**4.40**
o3-mini	OpenAI	Janvier 2025	**1.10**	**4.40**
o1	OpenAI	Décembre 2024	15.00	60.00
GPT-5	OpenAI	Aout 2025	1.25	10
Gemini 2.5 pro	Google	Mars 2025	1.25	10
Claude Opus 4	Anthropic	Mai 2025	15.00	75.00
Claude Sonnet 4	Anthropic	Mai 2025	3.00	15.00

Les modèles ont été invoqués dans leur version la plus récente (*snapshot*), sans configuration d'hyperparamètres spécifique, en s'appuyant sur les valeurs par défaut, sauf pour l'effort de raisonnement (*reasoning effort*) fixé à *high* afin de maximiser le niveau de réflexion interne des modèles. Les appels ont été effectués via le client Python de l'API OpenAI, en ciblant l'endpoint *v1/responses*, et en maintenant la séparation conventionnelle entre messages de type *system* et *user*.

4.5.3 PROMPT ENGINEERING

La qualité des règles générées par un modèle de raisonnement est intrinsèquement liée à la qualité des instructions qui lui sont fournies. En effet, aussi performant que soit le modèle sous-jacent, sa capacité à produire des résultats pertinents, cohérents et exploitables dans un contexte métier spécifique dépend en grande partie de la manière dont la tâche lui est formulée. C'est précisément l'objet du *prompt engineering*, qui constitue, dans notre protocole, une étape à part entière du processus de génération. Cette sous-section présente d'abord les fondements de cette discipline, puis détaille les techniques retenues et leur application concrète, avant de décrire la manière dont les règles produites ont été structurées et consolidées en vue de leur évaluation.

Définition et principes généraux

Afin de guider le raisonnement des modèles et d'en maximiser la pertinence dans le contexte de la génération de règles, la rédaction d'un prompt précis, structuré et explicite s'est avérée indispensable. Un prompt désigne l'instruction ou l'ensemble d'instructions fournies à un LLM pour orienter sa réponse. Le prompt engineering regroupe l'ensemble des techniques

visant à optimiser la formulation de ces instructions afin d'obtenir des résultats plus fiables, plus cohérents et mieux adaptés à la tâche visée. Elle entretient un lien étroit avec le Chain-of-Thought, dans la mesure où des prompts bien structurés permettent d'activer et de canaliser explicitement les capacités de raisonnement séquentiel des modèles. Étant donné la nature ouverte et créative de la tâche de génération de règles, la qualité du prompt constitue un levier déterminant pour la pertinence des sorties obtenues.

Techniques appliquées

Dans le cadre de cette recherche, quatre techniques d'ingénierie des prompts ont été appliquées de manière conjointe.

L'assignation de rôle (*role assignment*) consiste à attribuer au LLM un persona spécifique afin d'orienter son comportement, de focaliser son attention sur un domaine précis et de prévenir des réponses superficielles ou génériques. En appliquant cette technique, le modèle a été instruit de se comporter simultanément comme un expert en analyse de données et comme un expert en distribution électrique, ancrant ainsi ses réponses dans une double expertise pertinente au problème traité.

Le contextual prompting (ou injection de connaissances, *knowledge injection*) consiste à enrichir le prompt avec des informations contextuelles spécifiques à la tâche, incluant des éléments que le modèle n'a pas nécessairement rencontrés lors de son pré-entraînement. Dans notre cas, le prompt a incorporé : le contexte métier de l'organisation, les règles de base ayant servi à identifier les associations positives, un échantillon représentatif de paires d'associations positives et négatives, un échantillon des faux positifs détectés, un extrait des valeurs SHAP calculées lors des étapes précédentes, ainsi qu'une description détaillée

de chacune des variables du jeu de données. Cette richesse contextuelle vise à ancrer le raisonnement du modèle dans la réalité opérationnelle du réseau, réduisant ainsi le risque de généralisations non pertinentes.

Le *negative prompting* consiste à instruire explicitement le modèle sur ce qu'il doit éviter dans sa réponse. Il peut s'agir de prévenir des erreurs récurrentes observées dans les cycles précédents, d'interdire l'emploi d'un jargon technique inadapté, de proscrire certaines formulations ambiguës ou encore d'exclure des structures de règles jugées non exploitables. Dans notre protocole, cette technique a été employée pour délimiter précisément l'espace de génération acceptable, notamment en interdisant la production de règles redondantes avec celles déjà existantes ou reposant sur des variables dont l'influence avait été écartée lors des étapes d'analyse précédentes.

La rédaction des prompts étant par nature un processus itératif, la technique de *self-critique* a également été intégrée dans le prompt. Le modèle était ainsi instruit de produire, en complément de ses règles générées, une évaluation critique de sa propre réponse, d'identifier ses limites et de formuler les informations supplémentaires qui lui auraient permis d'améliorer ses résultats. À chaque cycle d'évaluation, les retours du *self-critique* étaient analysés et incorporés dans une nouvelle version du prompt, alimentant un processus d'amélioration continue, représenté schématiquement à la Figure 4.3.

Structure et format du prompt

Enfin, le formatage contraint (formatting constraint) a été appliqué à deux niveaux. D'une part, le prompt lui-même a été rédigé en syntaxe Markdown, structuré en plusieurs sections,

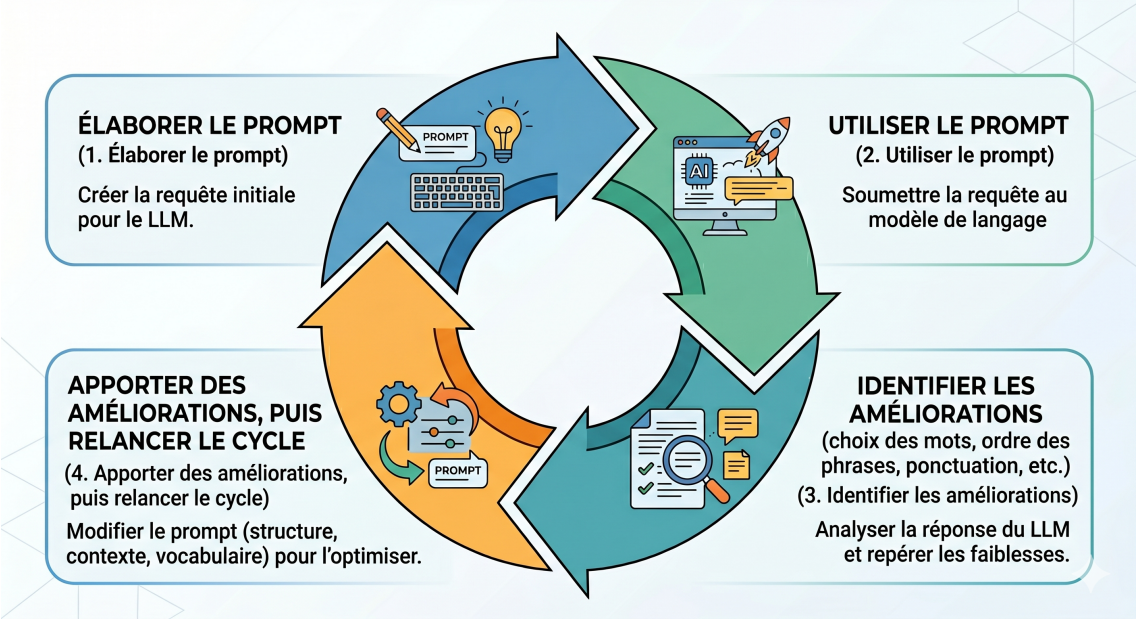

Figure 4.3 : Le processus itératif de rédaction d'un prompt

titres et listes afin d'en améliorer la lisibilité et la cohérence logique. D'autre part, le modèle a été instruit de fournir sa réponse sous la forme d'un objet JSON structuré comportant trois champs distincts : (1) une explication détaillée de son processus de création des règles, champ dont l'inclusion vise à garantir la transparence et la traçabilité du raisonnement suivi ; (2) les nouvelles règles générées ; et (3) le self-critique associé à la réponse. La structure finale du prompt est représentée schématiquement à la Figure 4.4. Le system prompt et le user prompt utilisés dans ce travail sont reproduits intégralement en Annexes A et B respectivement, afin de permettre au lecteur d'apprécier concrètement la mise en œuvre des principes de prompt engineering décrits ci-dessus.

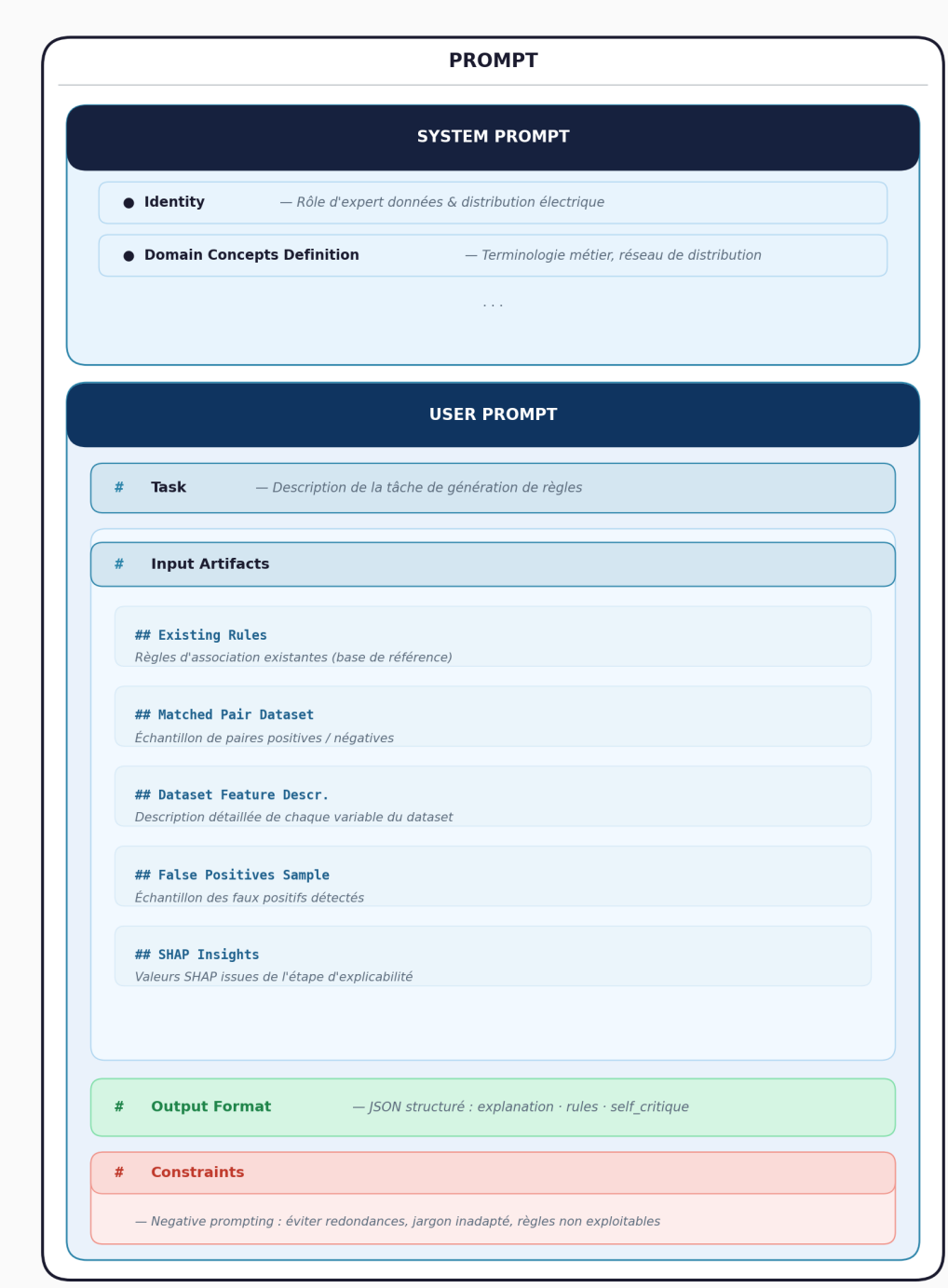

Figure 4.4 : Structure globale du prompt

Consolidation et stockage des règles produites

À l'issue de chaque cycle itératif, les règles générées par les modèles ont été extraites du champ JSON correspondant et consignées dans une table dédiée. Chaque entrée de cette table enregistre, d'une part, le texte intégral de la règle produite et, d'autre part, le modèle LLM ayant généré cette règle, qu'il s'agisse de o3-mini, o1 ou o4-mini, permettant ainsi une traçabilité complète de la provenance de chaque règle. Cette structuration facilite les comparaisons inter-modèles et prépare le terrain pour l'évaluation empirique de la pertinence des règles, décrite dans la section suivante.

CHAPITRE V

EXPÉRIMENTATIONS ET RÉSULTATS

La qualité des règles d'association générées par les modèles de langage étudiés ne peut être appréciée à travers un prisme unique d'analyse. En effet, l'évaluation de sorties issues de systèmes fondés sur des LLMs, d'une part, une mesure objective et reproductible des propriétés formelles et sémantiques des textes produits, et d'autre part, une appréciation contextuelle que seuls des évaluateurs compétents, humains ou assistés, sont en mesure de fournir. C'est précisément pour répondre à cette double exigence épistémique que le présent protocole d'évaluation s'articule autour de deux volets complémentaires et indissociables : une évaluation quantitative et une évaluation qualitative.

L'évaluation quantitative vise à produire des indicateurs chiffrés, comparables et stables, permettant de situer chaque modèle par rapport aux autres sur des axes de performance bien définis. Elle relève de ce que la littérature désigne sous le terme d'évaluation intrinsèque (*intrinsic evaluation*) : elle porte sur les propriétés internes et la qualité du texte généré en lui-même, indépendamment de son usage. Elle offre la rigueur nécessaire à toute démarche scientifique en garantissant la réplicabilité des résultats et en réduisant les biais liés à la subjectivité humaine. Toutefois, elle demeure insuffisante à elle seule, car aucune métrique

automatique ne peut, en l'état actuel de l'art, capturer intégralement la pertinence opérationnelle d'une nouvelle règle d'association dans un contexte d'optimisation.

L'évaluation qualitative vient donc en complément, en s'inscrivant davantage dans une logique d'évaluation extrinsèque (*extrinsic evaluation*) : elle évalue dans quelle mesure les règles générées s'avèrent effectivement utiles et applicables dans un contexte réel d'association des avis aux interruptions planifiées. Elle mobilise un panel d'experts métier pour juger de la cohérence, de la faisabilité et de l'utilité pratique des règles générées, et s'appuie également sur le paradigme émergent du *LLM-as-judge* pour obtenir une évaluation automatisée à large échelle reproduisant le raisonnement évaluatif humain. Cette double validation, quantitative et qualitative, garantit ainsi la robustesse et la validité des conclusions tirées de l'expérimentation.

5.1 ÉVALUATION QUANTITATIVE

L'évaluation quantitative repose sur deux axes complémentaires :un ensemble de métriques automatiques de qualité textuelle, et (ii) l'analyse de la réduction du déficit d'appariement, c'est-à-dire la comparaison, avant et après application des règles générées, du nombre d'avis et d'interruptions planifiées qui ne pouvaient jusqu'alors être associés à aucune règle existante. Ces deux volets sont présentés successivement ci-après.

5.1.1 MÉTRIQUES AUTOMATIQUES D'ÉVALUATION DE LA QUALITÉ TEXTUELLE

Les métriques automatiques constituent le premier levier d'objectivation de la qualité des règles générées. Elles permettent de comparer, de manière systématique et reproductible, les sorties des trois modèles étudiés par rapport à un ensemble de règles de référence établies par des

experts. Cinq métriques ont été retenues pour cette étude, sélectionnées afin de couvrir à la fois les dimensions lexicales, sémantiques et distributionnelles de la qualité textuelle. Ces métriques s'inscrivent toutes dans le cadre de l'évaluation intrinsèque, en ce qu'elles évaluent directement les propriétés internes du texte généré par comparaison à un corpus de référence, ici nos règles de base, et indirectement la qualité des LLMs. Nous nous sommes attardés néanmoins sur deux (2) familles fondamentalement distinctes par leur mode de fonctionnement : les métriques à base de n-grammes d'une part, et les métriques sémantiques d'autre part.

La première famille regroupe les métriques qui évaluent les textes en se fondant sur leur forme de surface, c'est-à-dire en comptabilisant les correspondances exactes de mots ou de séquences de mots entre le texte généré et le texte de référence. Ces approches reposent sur des statistiques de co-occurrence lexicale et présentent l'avantage d'être simples, rapides et interprétables. C'est dans ce cadre conceptuel que s'inscrivent les deux premières métriques retenues, BLEU [39] et ROUGE [33].

BLEU, acronyme de Bilingual Evaluation Understudy, est une métrique d'évaluation automatique initialement conçue pour la traduction automatique. Elle mesure la proximité entre un texte généré et des textes de référence en calculant un taux de chevauchement de n-grammes, c'est-à-dire en comptabilisant le nombre de séquences de mots communes entre la sortie du modèle et les références, tout en limitant ce décompte au nombre maximal d'occurrences observées dans une référence donnée. Dans le cadre de la présente étude, elle permet d'apprécier la fidélité lexicale des règles générées par rapport aux formulations de référence.

Si BLEU privilégie la précision, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) adopte la perspective inverse en mettant l'accent sur le rappel, c'est-à-dire en mesurant dans quelle proportion le contenu informatif des références se retrouve effectivement dans le texte

généré. Conçue à l'origine pour l'évaluation de résumés automatiques, elle se décline en plusieurs variantes complémentaires : ROUGE-N quantifie le chevauchement de n-grammes, ROUGE-L s'appuie sur la plus longue sous-séquence commune pour capturer l'ordre des mots à l'échelle de la phrase, ROUGE-W accorde un poids plus élevé aux correspondances spatiales consécutives, et ROUGE-S mesure les co-occurrences de bigrammes à distance. Cette richesse de variantes fait de ROUGE un outil particulièrement adapté à l'évaluation de la complétude des règles générées, quoique dans notre cas nous utiliserons uniquement ROUGE-L car elle évalue la similarité des séquences sans exiger de correspondances exactes de n-grammes consécutifs. Pour autant, elle partage avec BLEU la même limite fondamentale : elle demeure aveugle aux équivalences sémantiques exprimées dans un vocabulaire différent de celui de la référence, ce qui représente un obstacle non négligeable dans un domaine technique où la synonymie terminologique est courante.

C'est précisément pour dépasser cette limitation que la seconde famille de métriques, dite de similiarité sémantique ou contextuelle, a été employée dans ce protocole. Parce que les métriques de chevauchement de *n-gram* échouent souvent lorsqu'un texte généré est parfaitement valide mais utilise un vocabulaire entièrement différent de la référence, ces métriques simulent les évaluateurs humains. Plutôt que de s'appuyer sur la correspondance exacte de chaînes de caractères, ces métriques représentent les textes au moyen de plongements vectoriels contextualisés profonds, issus de modèles d'*embedding* tels que BERT [13]. Dans cet espace vectoriel continu, les mots de sens proche sont projetés à proximité les uns des autres, et le contexte d'utilisation de chaque terme est intégré dans sa représentation. Il en résulte une capacité à reconnaître naturellement les synonymes, les paraphrases et les dépendances à longue distance, permettant de récompenser correctement un texte généré qui exprime exactement le même sens que la référence, même s'il mobilise un vocabulaire ou des structures

grammaticales entièrement différents. De ce fait, les métriques sémantiques présentent une corrélation nettement plus élevée avec les jugements humains de qualité textuelle que les approches lexicales traditionnelles, ce qui justifie pleinement leur inclusion dans un protocole d'évaluation rigoureux. Une métrique de cette famille a été retenue : BERTScore [53]

BERTScore constitue le premier représentant de cette famille dans notre protocole. Son principe repose sur la projection des tokens du texte généré et du texte de référence dans un espace vectoriel contextuel produit par Bidirectional Encoder Representations from Transformers (BERT), suivie du calcul de la similarité cosinus entre ces représentations respectives. Un appariement glouton associe ensuite chaque token à son homologue le plus similaire dans l'autre texte, avec la possibilité d'intégrer une pondération par fréquence inverse de document afin d'ajuster l'importance relative des termes selon leur rareté dans le corpus. Cette capacité à capturer des équivalences au-delà des correspondances lexicales exactes revêt une importance particulière dans le domaine EPD (Electric Power Domain), où la même réalité technique peut être désignée par des formulations variées.

L'ensemble de ces trois (3) métriques forme ainsi un dispositif d'évaluation progressif et complémentaire, allant de la vérification lexicale de surface jusqu'à l'appréciation distributionnelle du sens. Afin d'en rendre lisibles les apports respectifs dans le contexte précis de cette étude, la Figure 5.1 présente les résultats comparatifs obtenus pour les trois modèles selon chacun de ces indicateurs. Cette vue synthétique permettra, plus tard, d'identifier les dimensions de qualité sur lesquelles les modèles se distinguent et celles sur lesquelles leurs performances convergent.

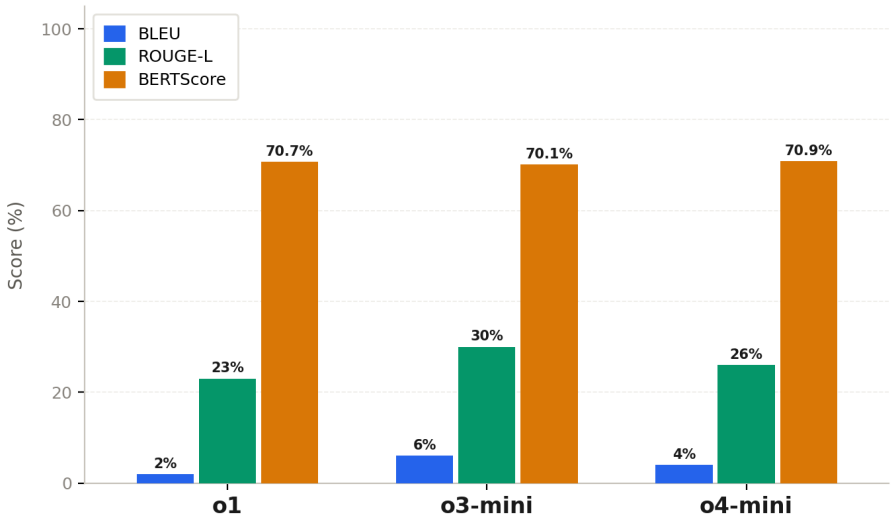

Figure 5.1 : Métriques d'évaluation automatiques pour l'ensemble des 3 modèles.

5.1.2 ANALYSE DE LA RÉDUCTION DU DÉFICIT D'APPARIEMENTS

Notre système de règles d'association ne tire sa valeur opérationnelle que dans la mesure où il parvient à établir des correspondances entre les entités du système, avis et interruptions planifiées, qui en étaient précédemment dépourvues. Or, dans notre corpus de données, une proportion non négligeable de ces entités demeure non appariée, faute de règles suffisamment généralisables pour les couvrir et occasionant un calcul défaillant Taux de Respect des Interruptions Planifiées (TRIP). C'est précisément ce déficit d'appariement que ce second volet de l'évaluation quantitative cherche à mesurer et à réduire.

Le principe de ce volet consiste à comparer le nombre d'avis et d'interruptions planifiées non appariés avant l'application des règles générées avec celui observé après leur intégration. Dans notre cas, une entité est considérée comme appariée dès lors qu'au moins une autre entité de nature différente lui est reliée, selon les règles préalablement définis. La différence entre ces deux états, avant et après, constitue le déficit d'appariement résiduel, dont la réduction

traduit directement la capacité du modèle à étendre la couverture des règles d'association à des situations jusqu'alors non prises en charge.

Il convient cependant de préciser les conditions dans lesquelles cette évaluation a pu être conduite. En raison de contraintes temporelles en cours au moment de la rédaction de ce mémoire, il n'a pas été possible de réaliser les calculs nécessaires à une comparaison exhaustive couvrant l'ensemble des années et des modèles étudiés. Les résultats disponibles se limitent à l'année 2025 et au modèle o4-mini, pour lequel le déficit d'appariement a néanmoins pu être mesuré sur l'ensemble des six (6) Centre d'Exploitation et de Distributions (CEDs) concernés. Bien que cette portée partielle ne permette pas de procéder à une comparaison inter-modèles complète telle qu'initialement envisagée, elle offre un premier aperçu empirique significatif de l'effet des règles générées sur la couverture associative des entités, et constitue une base solide pour des travaux d'évaluation ultérieurs à plus large échelle.

Cette mesure demeure directement interprétable dans un contexte opérationnel : une réduction du déficit d'appariement signifie qu'un plus grand nombre de situations fondées sur des règles peuvent désormais être considérées, fiabilisant ainsi concrètement le périmètre d'aide à la décision offert à la direction. Les résultats obtenus pour le modèle o4-mini sur l'année 2025, déclinés par CED, sont présentés dans le tableau 5.1 ci-contre.

5.2 ÉVALUATION QUALITATIVE

Si l'évaluation quantitative permet de situer objectivement les modèles les uns par rapport aux autres selon des critères mesurables, elle ne saurait suffire à elle seule à rendre compte de la valeur pratique des règles d'association générées. En effet, un corpus généré peut obtenir

Tableau 5.1 : Pourcentage de baisse des entités non-associées par CED après application des règles générées par le modèle o4-mini (année 2025)

	Baisse du nombre d'avis non-associés (%)	**Baisse du nombre d'interruptions non-associées (%)**
CED A	4,08	1,13
CED B	9,72	56,17
CED C	1,55	17,78
CED D	82,26	94,89
CED E	4,37	32,21
CED F	5,84	39,83

des scores élevés selon les métriques automatiques tout en s'avérant inapplicable dans le contexte opérationnel de l'appariement entre avis et interruptions, en raison de contraintes difficilement quantifiables. À l'inverse, il peut présenter de faibles performances automatiques tout en démontrant une pertinence pratique adéquate. C'est pour combler cette limitation structurelle que le protocole intègre un second volet, entièrement qualitatif, articulé autour de deux dispositifs complémentaires : la validation par un panel d'experts et le paradigme du *LLM-as-judge*.

5.2.1 VALIDATION PAR UN PANEL D'EXPERTS

La validation par des experts du domaine constitue une approche de référence en évaluation de systèmes d'intelligence artificielle appliqués à des contextes industriels ou professionnels. Elle repose sur le principe que la pertinence d'une règle d'association ne peut être pleinement appréciée qu'au regard des contraintes réelles de mise en œuvre, des pratiques établies du métier et du jugement professionnel accumulé au fil de l'expérience. Aucune métrique automatique, aussi sophistiquée soit-elle, ne peut se substituer à cette forme de connaissance tacite et contextualisée.

Dans le cadre de la présente étude, un panel composé de deux experts au sein des CEDs (Centre d'Exploitation et de Distribution) a été sollicité. Ces experts ont été sélectionnés en raison de leur connaissance approfondie de la conduite de réseau électrique et de leur capacité à évaluer de manière critique la cohérence et la faisabilité des règles proposées. Afin de structurer le processus d'évaluation et d'en garantir la comparabilité inter-évaluateurs, un formulaire d'évaluation standardisé a été élaboré sur Microsoft Forms et soumis à chacun d'entre eux par courriel.

Ce formulaire mobilise une échelle de Likert à cinq points, dont les modalités ont été formulées de manière à ancrer chaque niveau dans une position d'accord clairement définie. Contrairement à un simple item de notation (rating scale), l'échelle de Likert agrège plusieurs dimensions d'appréciation en un score composite, ce qui en renforce la fiabilité et la validité de mesure [5]. Ainsi, la valeur 1 correspond à « Pas du tout d'accord », traduisant un rejet explicite de l'énoncé évalué ; la valeur 2 correspond à « Pas d'accord », exprimant une position défavorable sans être catégorique ; la valeur 3 correspond à « Neutre », indiquant une absence de position tranchée ou une appréciation mitigée ; la valeur 4 correspond à « D'accord », signalant une adhésion générale à l'énoncé ; et la valeur 5 correspond à « Tout à fait d'accord », traduisant une approbation pleine et entière. Cette gradation symétrique en cinq niveaux permet de saisir avec nuance l'orientation de chaque expert à l'égard des dimensions évaluées, tout en produisant des données ordinales susceptibles d'être agrégées et comparées entre évaluateurs et entre modèles.

Les grandes dimensions retenues pour chaque règle soumise à évaluation sont : la qualité de la rédaction, l'exactitude et conformité aux données, le raisonnement, la transparence ainsi que la faisabilité opérationnelle. Un espace de commentaires libres est également prévu afin

de permettre aux experts de justifier leurs évaluations ou d'apporter des nuances non saisies par les items fermés. La structure complète du formulaire, telle qu'elle a été soumise aux participants, est présentée au Tableau 5.2, et les réponses au Tableau 5.3 (sans la section des commentaires). La version intégrale du formulaire telle qu'administrée est reproduite en Annexe C.

Tableau 5.2 : Structure du formulaire d'évaluation soumis aux experts

Dimension évaluée	Description	Échelle
Qualité de la documentation	La structure et la description des règles sont claires et exploitables	Likert 5 points
Exactitude et conformité aux données	Les règles générées reflètent la réalité technique du réseau électrique	Likert 5 points
Raisonnement et logique	Les règles générées reposent sur un enchaînement logique et justifié entre les éléments qu'elles associent	Likert 5 points
Impact opérationnel	Les règles générées sont applicables dans le contexte opérationnel réel et sont capables d'améliorer les indicateurs de performance	Likert 5 points
Transparence et éthique	Les règles sont équitables et transparentes pour être comprises facilement	Likert 5 points
Commentaires libres	Observations, nuances ou remarques complémentaires	Champ texte ouvert

Il convient de souligner que cette démarche a reçu l'approbation du Comité d'éthique de la recherche de l'Université du Québec à Chicoutimi (CER-UQAC), sous le certificat numéro 2026-2408, préalablement au démarrage de la phase de collecte des avis, afin de se soumettre aux exigences déontologiques encadrant la recherche universitaire impliquant des participants humains. Les experts ont été pleinement informés des objectifs de l'étude et de l'utilisation prévue de leurs évaluations. L'ensemble des avis collectés est traité de manière confidentielle et anonyme. À ce titre, le certificat d'approbation éthique est disponible en Annexe D.

Tableau 5.3 : Évaluation par les experts selon différentes dimensions

Dimension		Expert 1			Expert 2		
		o1	o3-mini	o4-mini	o1	o3-mini	o4-mini
Qualité de la documentation	Lisibilité et professionalisme	4	4	4	5	5	5
	cohérence de l'enchainement	4	4	4	5	5	5
	Accessibilité	4	4	4	5	5	5
Exactitude et conformité aux données	Fidélité aux référentiels métier	4	4	4	5	5	5
	Gestion des cas particuliers	3	4	3	3	2	2
Raisonnement et logique	Validité du raisonnement	3	3	3	2	2	2
Impact opérationnel	Potentiel d'efficacité	3	3	3	2	2	2
Transparence et éthique	Interprétabilité	3	2	3	3	3	2
	Equité	3	1	3	2	2	3

5.2.2 LE PARADIGME DU LLM-AS-JUDGE

En complément de la validation humaine, et dans le but de disposer d'une évaluation qualitative systématique, le protocole intègre le paradigme du *LLM-as-judge*. Ce cadre méthodologique repose sur l'utilisation d'un LLM plus performant distinct des modèles évalués en qualité de juge automatique, chargé d'apprécier la qualité des sorties générées selon des critères prédéfinis.

Il convient cependant de souligner que le périmètre d'évaluation confié au modèle juge a été délibérément circonscrit aux critères ne nécessitant ni accès aux données opérationnelles de l'entreprise, ni expertise métier spécialisée du réseau électrique de distribution. En effet, certains critères du formulaire tels que la fidélité aux référentiels métier, la gestion des cas particuliers et le potentiel d'efficacité, requièrent une connaissance approfondie des systèmes d'information internes et des réalités du terrain, qui excède les capacités d'un modèle de langue généraliste et auxquelles celui-ci ne doit pas avoir accès pour des raisons de confidentialité et de rigueur évaluative. Ces critères demeurent donc l'apanage exclusif du panel d'experts humains. En revanche, six critères ont été jugés accessibles et pertinents pour une évaluation automatisée : la lisibilité et le professionnalisme, la cohérence de l'enchaînement, l'accessibilité, la validité du raisonnement, l'interprétabilité et l'équité. Ces dimensions portent en effet sur des propriétés intrinsèques du texte généré, sa clarté, sa logique interne, sa neutralité et son intelligibilité, que le modèle juge est en mesure d'apprécier sans recourir à des informations contextuelles privilégiées.

Sur le plan opérationnel, l'évaluation par LLM-as-judge a été conduite via Microsoft Copilot, en mobilisant le modèle GPT-5.5 d'OpenAI. Pour chaque ensemble de règles soumis à évaluation, un prompt structuré et constant a été administré au modèle juge, garantissant ainsi l'homogénéité des conditions d'évaluation entre les différentes règles et les différents modèles comparés. Le prompt soumet la règle à évaluer selon les six critères retenus, en invitant le modèle juge à attribuer un score de 1 à 5 pour chacun d'eux, selon la même échelle de Likert utilisée par les experts humains.

La complémentarité entre le LLM-as-judge et la validation experte est ainsi structurée selon une logique de division des rôles évaluatifs : les experts humains couvrent l'ensemble des neuf

critères, incluant ceux exigeant une connaissance opérationnelle fine, tandis que le modèle juge intervient sur les six dimensions relevant de la qualité textuelle et logique intrinsèque des règles générées. Les scores attribués par le modèle juge pour chacun de ces critères, déclinés par modèle évalué, sont consignés dans le Tableau 5.4. La concordance entre ces deux sources d'évaluation sur les critères partagés constituera un indicateur de robustesse des conclusions.

Tableau 5.4 : Scores attribués par le LLM-as-judge par critère et par modèle évalué

Dimension		**LLM**		
		o1	o3-mini	o4-mini
Qualité de la documentation	Lisibilité et professionnalisme	4	4	4
	cohérence de l'enchaînement	3	3	3
	Accessibilité	4	4	3
Raisonnement et logique	Validité du raisonnement	3	3	2
Transparence et éthique	Interprétabilité	4	4	4
	Équité	4	4	3

5.3 SYNTHÈSE DU PROTOCOLE D'ÉVALUATION

Le protocole d'évaluation décrit dans cette section vise à fournir une appréciation multidimensionnelle et rigoureuse des règles d'association générées par les trois modèles o3-mini, o1 et o4-mini comparés. La combinaison de métriques automatiques (BLEU, ROUGE, BERTScore), du nombre de nouvelles entités isolées, d'une validation par un panel d'experts et d'une évaluation par LLM constituent un cadre d'évaluation à la fois robuste sur le plan scientifique et pertinent sur le plan applicatif. L'articulation entre évaluation intrinsèque et évaluation extrinsèque, entre mesure automatique et jugement humain, répond à l'exigence fondamentale de toute recherche appliquée en intelligence artificielle : celle de ne pas dissocier la performance

formelle des modèles de leur valeur d'usage réelle dans le contexte pour lequel ils ont été conçus. La triangulation des sources d'évaluation offre ainsi les garanties nécessaires à la validité de la conclusion présentée dans le chapitre suivant.

CHAPITRE VI

CONCLUSION

Ce travail exploratoire avait pour ambition de concevoir un cadre hybride combinant apprentissage automatique classique, IA explicable (XAI) et grands modèles de langage (LLMs) pour optimiser les règles d'association entre avis et interruptions planifiées dans le secteur de l'énergie chez Hydro-Québec. L'objectif ultime était de réduire le nombre d'entités non appariées et d'améliorer la fiabilité du calcul de l'indicateur Taux de Respect des Interruptions Plannifiées (TRIP). Pour en mesurer l'atteinte, les résultats ont été analysés selon deux axes complémentaires : quantitatif et qualitatif.

Sur le plan quantitatif, les résultats sont encourageants. Les scores BERTScore obtenus pour les trois modèles évalués (o1 : 70,7%, o3-mini : 70,1%, o4-mini : 70,9%) confirment que les règles générées par les LLMs entretiennent un lien sémantique solide avec les règles de base existantes. De plus, l'implémentation des règles générées par le modèle OpenAI o4-mini a permis de réduire significativement le nombre d'avis et d'interruptions non associés pour plusieurs Centre d'Exploitation et de Distributions (CEDs), augmentant ainsi la proportion d'entités intégrées dans le calcul du TRIP et, par conséquent, la fiabilité de cet indicateur décisionnel.

Ces résultats chiffrés, bien que prometteurs, ne suffisent cependant pas à eux seuls à valider la pertinence opérationnelle du système. C'est précisément ce que l'évaluation qualitative est venue nuancer. L'évaluation conduite à la fois par un LLM-as-judge et par des experts humains révèle un consensus satisfaisant sur la qualité générale de la documentation produite et sur la cohérence du raisonnement des modèles. Cependant, des divergences persistent, notamment sur les dimensions de transparence/équité et surtout d'impact opérationnel, cette dernière étant précisément celle que les experts métier sont les mieux positionnés pour juger. Ces derniers estiment que les règles générées manquent encore de pertinence opérationnelle suffisante pour une intégration directe en production, ce qui soulève des questions importantes quant à la maturité du système pour un usage terrain.

À la lumière de ces résultats, il est désormais possible de répondre à la question centrale de cette recherche : *Comment les LLMs, enrichis par les cadres d'IA explicable, peuvent-ils être exploités pour optimiser les règles d'association des avis aux interruptions planifiées dans le secteur de l'énergie ?* Cette étude démontre qu'un tel cadre hybride est techniquement faisable et sémantiquement pertinent : les LLMs sont capables de générer des règles d'association cohérentes avec les pratiques métier existantes et de contribuer à la réduction des entités isolées. Toutefois, si la preuve de concept est établie, les écarts relevés par les experts sur l'impact opérationnel rappellent que la faisabilité technique ne suffit pas ; une intégration opérationnelle complète requiert encore un travail d'alignement approfondi avec les exigences terrain.

Cet écart entre performance sémantique et adéquation opérationnelle s'explique en partie par les limites inhérentes à cette étude. Premièrement, la contrainte de temps n'a pas permis de calculer le nombre d'entités non appariées pour les modèles o1 et o3-mini, ce qui restreint la comparaison quantitative au seul modèle o4-mini et rend difficile toute conclusion comparative

entre modèles. Deuxièmement, l'étude s'est limitée aux modèles OpenAI, en raison d'un accès privilégié via un partenariat Microsoft. Des modèles de raisonnement d'autres fournisseurs leaders, Anthropic et Google, n'ont pu être testés, tout comme des modèles open-source prometteurs d'Alibaba ou NVIDIA, écartés pour des raisons de conformité interne et de capacité de calcul insuffisante pour un déploiement local. Cette restriction réduit la portée des conclusions et invite à la prudence quant à leur généralisation à l'ensemble du paysage des LLMs disponibles.

Ces limites, loin de disqualifier les apports de cette recherche, ouvrent au contraire des pistes de développement riches et structurantes. La principale avenue réside dans la transformation du pipeline actuel, linéaire et unidirectionnel, en un système agentique bouclé intégrant le principe de *human-in-the-loop*. Dans cette architecture repensée, les règles générées par les LLMs seraient soumises à l'inspection d'experts humains avant validation. Une fois validées, un agent IA prendrait en charge l'exécution autonome des étapes suivantes : formulation des requêtes SQL, transformations de données, réalisation des nouvelles associations et calcul automatique de la valeur actualisée du TRIP. Cette évolution vers l'IA agentique permettrait non seulement de renforcer la robustesse opérationnelle du système, mais aussi de répondre directement aux réserves exprimées par les experts en instaurant une validation humaine explicite à chaque cycle. À plus long terme, l'extension à d'autres modèles et fournisseurs, ainsi qu'une évaluation sur des données multi-CED plus larges, permettraient de consolider et de généraliser les conclusions de cette recherche, faisant de ce travail exploratoire le socle d'une solution industrielle viable.

BIBLIOGRAPHIE

BIBLIOGRAPHIE

[1] "Introducing Claude 4 — anthropic.com," https://www.anthropic.com/news/claude-4, [Accessed 12-06-2026].

[2] "Gemini 2.5 Pro | Gemini Enterprise Agent Platform | Google Cloud Documentation — docs.cloud.google.com," https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/2-5-pro, [Accessed 12-06-2026].

[3] "Introducing OpenAI o3 and o4-mini — openai.com," https://openai.com/index/introducing-o3-and-o4-mini/, [Accessed 12-06-2026].

[4] M. Abbasi, R. I. Nishat, C. Bond, J. B. Graham-Knight, P. Lasserre, Y. Lucet, et H. Najjaran, "A review of ai and machine learning contribution in predictive business process management (process enhancement and process improvement approaches)," *arXiv preprint arXiv :2407.11043*, 2024.

[5] J. Amidei, P. Piwek, et A. Willis, "The use of rating and likert scales in natural language generation human evaluation tasks : A review and some recommendations," ser. Proceedings of the 12th International Conference on Natural Language Generation. Association for Computational Linguistics, Conference Proceedings, p. 397–402. [En ligne]. Disponible : https://aclanthology.org/W19-8648/https://doi.org/10.18653/v1/W19-8648

[6] F. Amjad, T. Korotko, et A. Rosin, "Review of llms applications in electrical power energy systems," *IEEE Access*, 2025.

[7] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, et R. Benjamins, "Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, p. 82–115, 2020.

[8] N. Barlaug, "Lemon : explainable entity matching," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, n°. 8, p. 8171–8184, 2022.

[9] G. E. P. Box et G. M. Jenkins, *Time Series Analysis : Forecasting and Control*, 3rd éd. USA : Prentice Hall PTR, 1994.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, et A. Askell, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, p. 1877–1901, 2020.

[11] A. Celikyilmaz, E. Clark, et J. Gao, "Evaluation of text generation : A survey," *arXiv preprint arXiv :2006.14799*, 2020.

[12] S. Chen, S. Bateni, S. Grandhi, X. Li, C. Liu, et W. Yang, "Denas : automated rule generation by knowledge extraction from neural networks," p. 813–825, 2020. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.1145/3368089.3409733

[13] J. Devlin, M.-W. Chang, K. Lee, et K. Toutanova, "Bert : Pre-training of deep bidirectional transformers for language understanding," 2019. [En ligne]. Disponible : https://arxiv.org/abs/1810.04805

[14] V. Di Cicco, D. Firmani, N. Koudas, P. Merialdo, et D. Srivastava, "Interpreting deep learning models for entity resolution : an experience report using lime," dans *Proceedings of the second international workshop on exploiting artificial intelligence techniques for data management*, Conference Proceedings, p. 1–4.

[15] A. Doan, A. Ardalan, J. Ballard, S. Das, Y. Govind, P. Konda, H. Li, S. Mudgal, E. Paulson, et G. P. Suganthan, "Human-in-the-loop challenges for entity matching : A midterm report," dans *Proceedings of the 2nd workshop on human-in-the-loop data analytics*, Conference Proceedings, p. 1–6.

[16] M. Dumas, M. La Rosa, J. Mendling, H. A. Reijers *et al.*, *Fundamentals of business process management*. Springer, 2013, vol. 1.

[17] A. Ebaid, S. Thirumuruganathan, W. G. Aref, A. Elmagarmid, et M. Ouzzani, "Explainer : Entity resolution explanations," dans *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, Conference Proceedings, p. 2000–2003.

[18] M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, et N. Tang, "Distributed representations of tuples for entity resolution," *Proc. VLDB Endow.*, vol. 11, n°. 11, p. 1454–1467, Jul. 2018. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.14778/3236187.3269461

[19] P. Fettke et C. Di Francescomarino, "Business process management and artificial intelligence," 2025.

[20] S. Ge, Y. Sun, Y. Cui, et D. Wei, "An innovative solution to design problems : Applying the chain-of-thought technique to integrate llm-based agents with concept generation methods," *IEEE Access*, vol. 13, p. 10 499–10 512, 2025.

[21] O. Golovneva, M. Chen, S. Poff, M. Corredor, L. Zettlemoyer, M. Fazel-Zarandi, et A. Celikyilmaz, "Roscoe : A suite of metrics for scoring step-by-step reasoning," *arXiv preprint arXiv :2212.07919*, 2022, https ://github.com/facebookresearch/ParlAI/tree/main/projects/roscoe.

[22] M. Grohs, L. Abb, N. Elsayed, et J.-R. Rehse, "Large language models can accomplish business process management tasks," dans *International conference on business process management*. Springer, Conference Proceedings, p. 453–465.

[23] S. Hao, Y. Gu, H. Ma, J. Hong, Z. Wang, D. Wang, et Z. Hu, "Reasoning with language model is planning with world model," dans *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Conference Proceedings, p. 8154–8173.

[24] S. Hochreiter et J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, n°. 8, p. 1735–1780, Nov. 1997. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.1162/neco.1997.9.8.1735

[25] M. Jin, S. Wang, L. Ma, Z. Chu, J. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan *et al.*, "Time-llm : Time series forecasting by reprogramming large language models," dans *International conference on learning representations*, vol. 2024, 2024, p. 23 857–23 880.

[26] S. Kaltenpoth et O. Müller, "Don't touch the power line - a proof-of-concept for aligned llm-based assistance systems to support the maintenance in the electricity distribution system," *SIGENERGY Energy Inform. Rev.*, vol. 4, n°. 4, p. 16–22, Feb. 2025. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.1145/3717413.3717415

[27] P. V. Konda, *Magellan : Toward building entity matching management systems*. The University of Wisconsin-Madison, 2018.

[28] H. Kourani, A. Berti, D. Schuster, et W. M. van der Aalst, "Process modeling with large language models," dans *International Conference on Business Process Modeling, Development and Support*. Springer, Conference Proceedings, p. 229–244.

[29] H. Kourani, A. Berti, J. Hennrich, W. Kratsch, R. Weidlich, C.-Y. Li, A. Arslan, W. M. van der Aalst, et D. Schuster, "Leveraging large language models for enhanced process model comprehension," *Decision Support Systems*, p. 114563, 2025.

[30] H. Kourani, A. Berti, D. Schuster, et W. M. van der Aalst, "Evaluating large language models on business process modeling : framework, benchmark, and self-improvement analysis : H. kourani et al," *Software and Systems Modeling*, p. 1–36, 2025.

[31] M. Legris, B. Bouchard, A. J. N. Nzeko'o, K. Bouchard, et M. Aoun-Allah, "Towards agile and transparent data management : New generative ai-driven tools

for data governance," dans *Proceedings of the 2025 International Conference on Information Technology for Social Good*, ser. GoodIT '25. New York, NY, USA : Association for Computing Machinery, 2025, p. 15–23. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.1145/3748699.3749768

[32] Y. Li, J. Li, Y. Suhara, A. Doan, et W.-C. Tan, "Deep entity matching with pre-trained language models," *Proc. VLDB Endow.*, vol. 14, n°. 1, p. 50–60, 2020, dITTO, domain knowledge injection, data augmentation, end-to-end GEM frameworksupport for different data schema. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.14778/3421424.3421431

[33] C.-Y. Lin, "Rouge : A package for automatic evaluation of summaries," ser. Text Summarization Branches Out. Association for Computational Linguistics, Conference Proceedings, p. 74–81. [En ligne]. Disponible : https://aclanthology.org/W04-1013/

[34] S. M. Lundberg et S.-I. Lee, "A unified approach to interpreting model predictions," p. 4768–4777, 2017.

[35] T. Mikolov, K. Chen, G. Corrado, et J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv :1301.3781*, 2013.

[36] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, et V. Raghavendra, "Deep learning for entity matching : A design space exploration," dans *Proceedings of the 2018 International Conference on Management of Data*, ser. SIGMOD '18. New York, NY, USA : Association for Computing Machinery, 2018, p. 19–34. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.1145/3183713.3196926

[37] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, et A. Mian, "A comprehensive overview of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 16, n°. 5, Aug. 2025. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.1145/3744746

[38] OpenAI, :, A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A. T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Zhang, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C. M. Li, C. de Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F. P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. von Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc,

H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H. W. Chung, I. Kivlichan, I. O'Connell, I. Osband, I. C. Gilaberte, I. Akkaya, I. Kostrikov, I. Sutskever, I. Kofman, J. Pachocki, J. Lennon, J. Wei, J. Harb, J. Twore, J. Feng, J. Yu, J. Weng, J. Tang, J. Yu, J. Q. Candela, J. Palermo, J. Parish, J. Heidecke, J. Hallman, J. Rizzo, J. Gordon, J. Uesato, J. Ward, J. Huizinga, J. Wang, K. Chen, K. Xiao, K. Singhal, K. Nguyen, K. Cobbe, K. Shi, K. Wood, K. Rimbach, K. Gu-Lemberg, K. Liu, K. Lu, K. Stone, K. Yu, L. Ahmad, L. Yang, L. Liu, L. Maksin, L. Ho, L. Fedus, L. Weng, L. Li, L. McCallum, L. Held, L. Kuhn, L. Kondraciuk, L. Kaiser, L. Metz, M. Boyd, M. Trebacz, M. Joglekar, M. Chen, M. Tintor, M. Meyer, M. Jones, M. Kaufer, M. Schwarzer, M. Shah, M. Yatbaz, M. Y. Guan, M. Xu, M. Yan, M. Glaese, M. Chen, M. Lampe, M. Malek, M. Wang, M. Fradin, M. McClay, M. Pavlov, M. Wang, M. Wang, M. Murati, M. Bavarian, M. Rohaninejad, N. McAleese, N. Chowdhury, N. Chowdhury, N. Ryder, N. Tezak, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, P. Chao, P. Ashbourne, P. Izmailov, P. Zhokhov, R. Dias, R. Arora, R. Lin, R. G. Lopes, R. Gaon, R. Miyara, R. Leike, R. Hwang, R. Garg, R. Brown, R. James, R. Shu, R. Cheu, R. Greene, S. Jain, S. Altman, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Hernandez, S. Baker, S. McKinney, S. Yan, S. Zhao, S. Hu, S. Santurkar, S. R. Chaudhuri, S. Zhang, S. Fu, S. Papay, S. Lin, S. Balaji, S. Sanjeev, S. Sidor, T. Broda, A. Clark, T. Wang, T. Gordon, T. Sanders, T. Patwardhan, T. Sottiaux, T. Degry, T. Dimson, T. Zheng, T. Garipov, T. Stasi, T. Bansal, T. Creech, T. Peterson, T. Eloundou, V. Qi, V. Kosaraju, V. Monaco, V. Pong, V. Fomenko, W. Zheng, W. Zhou, W. Zhan, W. McCabe, W. Zaremba, Y. Dubois, Y. Lu, Y. Chen, Y. Cha, Y. Bai, Y. He, Y. Zhang, Y. Wang, Z. Shao, et Z. Li, "Openai o1 system card," 2026. [En ligne]. Disponible : https://arxiv.org/abs/2412.16720

[39] K. Papineni, S. Roukos, T. Ward, et W.-J. Zhu, "Bleu : a method for automatic evaluation of machine translation," p. 311–318, 2002. [En ligne]. Disponible : https://doi.org/10.3115/1073083.1073135

[40] J. Pennington, R. Socher, et C. Manning, "GloVe : Global vectors for word representation," dans *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, et W. Daelemans, édit. Doha, Qatar : Association for Computational Linguistics, Oct. 2014, p. 1532–1543. [En ligne]. Disponible : https://aclanthology.org/D14-1162/

[41] M. A. Rastaghi, E. Kamalloo, et D. Rafiei, "Probing the robustness of pre-trained language models for entity matching," p. 3786–3790, 2022. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.1145/3511808.3557673

[42] M. T. Ribeiro, S. Singh, et C. Guestrin, ""why should i trust you?" : Explaining the predictions of any classifier," p. 1135–1144, 2016. [En ligne]. Disponible : https://doi.org/10.1145/2939672.2939778

[43] N. Shahbazi, N. Danevski, F. Nargesian, A. Asudeh, et D. Srivastava, “Through the fairness lens : Experimental analysis and evaluation of entity matching,” *Proc. VLDB Endow.*, vol. 16, n°. 11, p. 3279–3292, 2023, voir ref 32, 44, 45, 53, 56. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.14778/3611479.3611525

[44] A. Singh, A. Fry, A. Perelman, A. Tart, A. Ganesh, A. El-Kishky, A. McLaughlin, A. Low, A. Ostrow, A. Ananthram, A. Nathan, A. Luo, A. Helyar, A. Madry, A. Efremov, A. Spyra, A. Baker-Whitcomb, A. Beutel, A. Karpenko, A. Makelov, A. Neitz, A. Wei, A. Barr, A. Kirchmeyer, A. Ivanov, A. Christakis, A. Gillespie, A. Tam, A. Bennett, A. Wan, A. Huang, A. M. Sandjideh, A. Yang, A. Kumar, A. Saraiva, A. Vallone, A. Gheorghe, A. G. Garcia, A. Braunstein, A. Liu, A. Schmidt, A. Mereskin, A. Mishchenko, A. Applebaum, A. Rogerson, A. Rajan, A. Wei, A. Kotha, A. Srivastava, A. Agrawal, A. Vijayvergiya, A. Tyra, A. Nair, A. Nayak, B. Eggers, B. Ji, B. Hoover, B. Chen, B. Chen, B. Barak, B. Minaiev, B. Hao, B. Baker, B. Lightcap, B. McKinzie, B. Wang, B. Quinn, B. Fioca, B. Hsu, B. Yang, B. Yu, B. Zhang, B. Brenner, C. R. Zetino, C. Raymond, C. Lugaresi, C. Paz, C. Hudson, C. Whitney, C. Li, C. Chen, C. Cole, C. Voss, C. Ding, C. Shen, C. Huang, C. Colby, C. Hallacy, C. Koch, C. Lu, C. Kaplan, C. Kim, C. Minott-Henriques, C. Frey, C. Yu, C. Czarnecki, C. Reid, C. Wei, C. Decareaux, C. Scheau, C. Zhang, C. Forbes, D. Tang, D. Goldberg, D. Roberts, D. Palmie, D. Kappler, D. Levine, D. Wright, D. Leo, D. Lin, D. Robinson, D. Grabb, D. Chen, D. Lim, D. Salama, D. Bhattacharjee, D. Tsipras, D. Li, D. Yu, D. Strouse, D. Williams, D. Hunn, E. Bayes, E. Arbus, E. Akyurek, E. Y. Le, E. Widmann, E. Yani, E. Proehl, E. Sert, E. Cheung, E. Schwartz, E. Han, E. Jiang, E. Mitchell, E. Sigler, E. Wallace, E. Ritter, E. Kavanaugh, E. Mays, E. Nikishin, F. Li, F. P. Such, F. de Avila Belbute Peres, F. Raso, F. Bekerman, F. Tsimpourlas, F. Chantzis, F. Song, F. Zhang, G. Raila, G. McGrath, G. Briggs, G. Yang, G. Parascandolo, G. Chabot, G. Kim, G. Zhao, G. Valiant, G. Leclerc, H. Salman, H. Wang, H. Sheng, H. Jiang, H. Wang, H. Jin, H. Sikchi, H. Schmidt, H. Aspegren, H. Chen, H. Qiu, H. Lightman, I. Covert, I. Kivlichan, I. Silber, I. Sohl, I. Hammoud, I. Clavera, I. Lan, I. Akkaya, I. Kostrikov, I. Kofman, I. Etinger, I. Singal, J. Hehir, J. Huh, J. Pan, J. Wilczynski, J. Pachocki, J. Lee, J. Quinn, J. Kiros, J. Kalra, J. Samaroo, J. Wang, J. Wolfe, J. Chen, J. Wang, J. Harb, J. Han, J. Wang, J. Zhao, J. Chen, J. Yang, J. Tworek, J. Chand, J. Landon, J. Liang, J. Lin, J. Liu, J. Wang, J. Tang, J. Yin, J. Jang, J. Morris, J. Flynn, J. Ferstad, J. Heidecke, J. Fishbein, J. Hallman, J. Grant, J. Chien, J. Gordon, J. Park, J. Liss, J. Kraaijeveld, J. Guay, J. Mo, J. Lawson, J. McGrath, J. Vendrow, J. Jiao, J. Lee, J. Steele, J. Wang, J. Mao, K. Chen, K. Hayashi, K. Xiao, K. Salahi, K. Wu, K. Sekhri, K. Sharma, K. Singhal, K. Li, K. Nguyen, K. Gu-Lemberg, K. King, K. Liu, K. Stone, K. Yu, K. Ying, K. Georgiev, K. Lim, K. Tirumala, K. Miller, L. Ahmad, L. Lv, L. Clare, L. Fauconnet, L. Itow, L. Yang, L. Romaniuk, L. Anise, L. Byron, L. Pathak, L. Maksin, L. Lo, L. Ho, L. Jing, L. Wu, L. Xiong, L. Mamitsuka, L. Yang, L. McCallum, L. Held, L. Bourgeois, L. Engstrom, L. Kuhn, L. Feuvrier, L. Zhang, L. Switzer, L. Kondraciuk, L. Kaiser, M. Joglekar, M. Singh, M. Shah, M. Stratta, M. Williams,

M. Chen, M. Sun, M. Cayton, M. Li, M. Zhang, M. Aljubeh, M. Nichols, M. Haines, M. Schwarzer, M. Gupta, M. Shah, M. Y. Guan, M. Huang, M. Dong, M. Wang, M. Glaese, M. Carroll, M. Lampe, M. Malek, M. Sharman, M. Zhang, M. Wang, M. Pokrass, M. Florian, M. Pavlov, M. Wang, M. Chen, M. Wang, M. Feng, M. Bavarian, M. Lin, M. Abdool, M. Rohaninejad, N. Soto, N. Staudacher, N. LaFontaine, N. Marwell, N. Liu, N. Preston, N. Turley, N. Ansman, N. Blades, N. Pancha, N. Mikhaylin, N. Felix, N. Handa, N. Rai, N. Keskar, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, O. Gleeson, P. Mishkin, P. Lesiewicz, P. Baltescu, P. Belov, P. Zhokhov, P. Pronin, P. Guo, P. Thacker, Q. Liu, Q. Yuan, Q. Liu, R. Dias, R. Puckett, R. Arora, R. T. Mullapudi, R. Gaon, R. Miyara, R. Song, R. Aggarwal, R. Marsan, R. Yemiru, R. Xiong, R. Kshirsagar, R. Nuttall, R. Tsiupa, R. Eldan, R. Wang, R. James, R. Ziv, R. Shu, R. Nigmatullin, S. Jain, S. Talaie, S. Altman, S. Arnesen, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Yoo, S. Heon, S. Ethersmith, S. Grove, S. Taylor, S. Bubeck, S. Banesiu, S. Amdo, S. Zhao, S. Wu, S. Santurkar, S. Zhao, S. R. Chaudhuri, S. Krishnaswamy, Shuaiqi, Xia, S. Cheng, S. Anadkat, S. P. Fishman, S. Tobin, S. Fu, S. Jain, S. Mei, S. Egoian, S. Kim, S. Golden, S. Mah, S. Lin, S. Imm, S. Sharpe, S. Yadlowsky, S. Choudhry, S. Eum, S. Sanjeev, T. Khan, T. Stramer, T. Wang, T. Xin, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Degry, T. Shadwell, T. Fu, T. Gao, T. Garipov, T. Sriskandarajah, T. Sherbakov, T. Korbak, T. Kaftan, T. Hiratsuka, T. Wang, T. Song, T. Zhao, T. Peterson, V. Kharitonov, V. Chernova, V. Kosaraju, V. Kuo, V. Pong, V. Verma, V. Petrov, W. Jiang, W. Zhang, W. Zhou, W. Xie, W. Zhan, W. McCabe, W. DePue, W. Ellsworth, W. Bain, W. Thompson, X. Chen, X. Qi, X. Xiang, X. Shi, Y. Dubois, Y. Yu, Y. Khakbaz, Y. Wu, Y. Qian, Y. T. Lee, Y. Chen, Y. Zhang, Y. Xiong, Y. Tian, Y. Cha, Y. Bai, Y. Yang, Y. Yuan, Y. Li, Y. Zhang, Y. Yang, Y. Jin, Y. Jiang, Y. Wang, Y. Wang, Y. Liu, Z. Stubenvoll, Z. Dou, Z. Wu, et Z. Wang, "Openai gpt-5 system card," 2026. [En ligne]. Disponible : https://arxiv.org/abs/2601.03267

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, et I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[46] J. Wang, Y. Li, W. Hirota, et E. Kandogan, "Machop : an end-to-end generalized entity matching framework," p. Article 2, 2022, megagon. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.1145/3533702.3534910

[47] P. Wang, X. Zeng, L. Chen, F. Ye, Y. Mao, J. Zhu, et Y. Gao, "Promptem : prompt-tuning for low-resource generalized entity matching," *Proc. VLDB Endow.*, vol. 16, n°. 2, p. 369–378, 2022, gEM, little labelled data, heterogeneous format. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.14778/3565816.3565836

[48] Z. Wang, C. Huang, et X. Yao, "A roadmap of explainable artificial intelligence : Explain to whom, when, what and how ?" *ACM Trans. Auton. Adapt. Syst.*, vol. 19, n°. 4, p. Article 20, 2024. [En ligne]. Disponible : https://doi-org.sbiproxy.uqac.ca/10.1145/3702004

[49] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, et D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, p. 24 824–24 837, 2022.

[50] R. Wu, S. Chaba, S. Sawlani, X. Chu, et S. Thirumuruganathan, "Zeroer : Entity resolution using zero labeled examples," dans *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, Conference Proceedings, p. 1149–1164.

[51] J. Xiao, W. Dut, Z. Xu, et Y. Qian, "Cross-system data integration based on rule-based nlp and node2vec," dans *2023 8th International Conference on Data Science in Cyberspace (DSC)*, Conference Proceedings, p. 566–570.

[52] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, et Y. Cao, "React : Synergizing reasoning and acting in language models," dans *The eleventh international conference on learning representations*, Conference Proceedings.

[53] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, et Y. Artzi, "Bertscore : Evaluating text generation with bert," *arXiv preprint arXiv :1904.09675*, 2019.

ANNEXE A : SYSTEM PROMPT

#IDENTITY

You are an expert in deterministic record linkage working in the Electric Power Domain. You reason carefully from statistical evidence before formalizing any rule.

You are developing a matching rule system for planned interruptions and notices. The system currently uses a set of rules written in French to form match pairs. These rules are insufficient causing a substantial population of interruptions and notices unmatched.

DOMAIN CONCEPTS

1. Planned Interruption :

An interruption in electrical service created to allow work to be performed while ensuring the safety of the personnel assigned to that work.

2. Centre d'exploitation de distribution (CED)

Geographical-based location where the distribution network is operated. There are currently 6 CEDs in Quebec province.

3. Notice :

A communication sent to customers to inform them of an interruption that is scheduled to occur.

4. Point

A point is a location where a customer is connected to the distribution network. A point refers here to a transformer.

5. Match

A match is a pair made up of a planned interruption and a notice that are considered to be part of a same work performed previously. The match is formed when the interruption and the notice are associated following some rules.

6. No-match

A no-match is a pair made up of a planned interruption and a notice that are not considered to be part of a same work performed previously. The no-match is formed applying negative sampling randomly on lonely items.

7. Lonely item

A lonely item is a notice that has not been matched to any planned interruption after following the existing rules and vice versa.

8. Negative sampling

Negative sampling is a technique used to form no-match pairs. It is a random sampling of pairs of items that are not matched.

ANNEXE B : USER PROMPT

TASK

Your task is to derive one or more new matching rules that will reduce the number of lonely items by generalising a latent matching pattern that the classifier has learn but has not yet been expressed as an explicit deterministic criterion.

INPUT ARTEFACTS

1. Existing Rules

These are the rules written in French and currently used in production.

‘‘‘

[REDACTED]

‘‘‘

2. Matched-pair dataset

Here is a sample dataset of pairs from a larger dataset formed by the existing rules. This dataset is used to train a classifier that predicts whether a pair notice-planned interruption is a match or no-match.

```

[REDACTED]

```

3. Dataset metadata

This is a description of the features used in the dataset.

```

[REDACTED]

```

4. False positives

Sample candidate pairs where the classifier model predicted a match but are labeled as no-match in the ground truth. These represent some pairs that should match and the existing rules have failed to capture.

```

[REDACTED]

```

5. SHAP insights

SHAP values outputed for the false positives sampled from the dataset.

```

[REDACTED]
```

‘‘‘

STEPS

Work through the following steps to derive new rules :

1. Understand the existing rules logic

2. Profile the false-positives population

3. Interpret the SHAP insights

4. Hypothesize the lattent pattern

5. Formalize your hypothesis as new rules

OUTPUT FORMAT

Provide your output written as a JSON string object with the following fields :

- ‘new_rules‘ : the proposed new rules

- ‘self_critique‘ : a non-technical explanation of how the new rule is derived from the existing rules

- ‘research_notes‘ : any additional notes about provided artifacts for future improvements

CONSTRAINTS

- Do not invent features not present in the dataset schema

- Ouput should be written in French

- Do not speculate about data outside the provided artifacts

- Do not explicitly mention the column names ; instead, refer directly to their meanings

- Introduce additional features from the dataset in the new rules

ANNEXE C : FORMULAIRE D'ÉVALUATION COMPLET

Ce formulaire a été administré via Microsoft Forms à chaque expert participant à l'évaluation qualitative des règles d'association générées. L'échelle de réponse utilisée pour les items 1 à 9 est une échelle de Likert à cinq points : 1 — Pas du tout d'accord, 2 — Pas d'accord, 3 — Neutre, 4 — D'accord, 5 — Tout à fait d'accord.

Section 1 — Qualité de la documentation

Objectif : La structure et la description du processus sont-elles claires et exploitables ?

#	Critère	Énoncé	1	2	3	4	5
1	Lisibilité et professionnalisme	Le texte est clair, correctement structuré et prêt à être diffusé sans correction linguistique.	○	○	○	○	○
2	Cohérence de l'enchaînement	La séquence des étapes respecte au moins la logique opérationnelle actuelle du terrain (Horodatage → Localisation CED → Transformateur commun).	○	○	○	○	○
3	Accessibilité	Les nouvelles règles sont compréhensibles pour un non-expert sans nécessiter une connaissance technique approfondie des données.	○	○	○	○	○

Section 2 — Exactitude et conformité aux données

Objectif : Les règles reflètent-elles la réalité technique du réseau électrique ?

#	Critère	Énoncé	1	2	3	4	5
4	Fidélité aux référentiels métier	Les nouvelles règles utilisent exclusivement des informations existantes dans nos systèmes (absence d'informations fictives ou erronées).	○	○	○	○	○
5	Gestion des cas particuliers	Les nouvelles règles traitent les scénarios imprévisibles et non seulement les cas de correspondance évidents.	○	○	○	○	○

Section 3 — Raisonnement et logique

Objectif : La pertinence des règles d'association.

#	Critère	Énoncé	1	2	3	4	5
6	Validité du raisonnement	Chaque nouveau critère de décision est techniquement justifié pour valider une association.	○	○	○	○	○

Section 4 — Impact opérationnel

Objectif : Capacité des règles à améliorer les indicateurs de performance.

#	Critère	Énoncé	1	2	3	4	5
7	Potentiel d'efficacité	L'adoption totale ou partielle des nouveaux critères mentionnés dans ces règles permettrait d'augmenter le nombre d'associations ou d'améliorer le taux de respect des interruptions planifiées.	○	○	○	○	○

Section 5 — Transparence et éthique

Objectif : Les règles sont-elles équitables et transparentes ?

#	Critère	Énoncé	1	2	3	4	5
8	Interprétabilité	Je comprends pourquoi le modèle suggère d'effectuer des associations sur la base de certains nouveaux critères.	○	○	○	○	○
9	Équité	Les nouveaux critères d'association sont exempts de biais (envers un secteur géographique, une catégorie de clients, etc.).	○	○	○	○	○

Section 6 — Conclusion de l'expert

10. Commentaires et suggestions d'amélioration

En cas de note inférieure ou égale à 3, ou pour toute autre raison, merci de préciser les points de blocage (ex. : étape illogique, champ de donnée manquant, règle trop permissive).

Note. Les cercles (○) représentent les boutons radio tels qu'ils apparaissaient dans le formulaire Microsoft Forms administré aux participants. L'échelle complète est : 1 — Pas du tout d'accord ; 2 — Pas d'accord ; 3 — Neutre ; 4 — D'accord ; 5 — Tout à fait d'accord.

ANNEXE D : CERTIFICAT D'APPROBATION ÉTHIQUE

Ce mémoire a fait l'objet d'une certification éthique auprès du CER-UQAC. Le numéro du certificat est 2026-240.