UQAC
Université du Québec
à Chicoutimi

CADRE MULTI-DIMENSIONNEL D'ÉVALUATION DES MÉTHODES D'INTELLIGENCE ARTIFICIELLE EXPLICABLE POUR LA DÉTECTION DE FRAUDE SUR BLOCKCHAINS PUBLIQUES

PAR DAVID LE SAGE OUMBE FOKOU

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

ABSTRACT

Explanations produced by Explainable Artificial Intelligence (XAI) methods are increasingly required to audit fraud detection systems deployed on public blockchains, yet most studies still evaluate them through proxy classification metrics such as F1 and AUC. This thesis addresses the gap by introducing a reproducible, multi-dimensional evaluation framework that jointly scores explanations along four complementary axes : fidelity, stability, domain alignment through a new Blockchain Rule Alignment Score (BRAS), and downstream usefulness assessed with a triad of Large Language Model (LLM) agents and a Machine Learning baseline. The framework is motivated by the four structural gaps identified by Zafar and Wu in their recent systematic review of 49 XAI studies on fraud detection, namely the prevalence of predictive surrogate metrics, the data-centric distortions induced by synthetic oversampling, the scarcity of human-centred evaluation, and the methodological monoculture around SHAP.

We instantiate the framework on the Elliptic Bitcoin and Ethereum Fraud datasets, across five explainers (SHAP, LIME, Integrated Gradients, GNNExplainer, GraphLIME) and four classifiers (Random Forest, LightGBM, Temporal Graph Convolutional Network, GraphSAGE). SHAP dominates on tabular models while Integrated Gradients wins on graph models. Adding explanations to LLM prompts raises Decision Accuracy from 0.52–0.79 to 0.81–0.94 and inter-agent Cohen's Kappa from near zero to near one. Once explanations are provided, the three LLM agents reach near-perfect agreement with a Random Forest trained on the same attributions, which suggests, within the limits of the evaluated sample (five instances per condition), that XAI outputs transfer the discriminative signal of the teacher model. The thesis further extends the conference paper with an expanded literature review, an in-depth methodological formalisation, complementary discussions on the orthogonality of the four axes, and a reflection on the operational implications for compliance pipelines.

Keywords : Explainable Artificial Intelligence, Blockchain, Fraud Detection, Graph Neural Networks, Large Language Models, Evaluation Framework, BRAS.

RÉSUMÉ

Les explications produites par les méthodes d'intelligence artificielle explicable (XAI) sont de plus en plus requises pour auditer les systèmes de détection de fraude déployés sur les blockchains publiques. Pourtant, la majorité des études évaluent encore ces explications au moyen de métriques de classification comme le score F1 ou l'AUC, qui ne renseignent pas directement sur la qualité explicative. Ce mémoire propose un cadre d'évaluation reproductible et multi-dimensionnel qui note conjointement les explications selon quatre axes complémentaires : la fidélité, la stabilité, l'alignement avec le domaine au moyen d'une nouvelle métrique appelée *Blockchain Rule Alignment Score* (BRAS), et l'utilité en aval mesurée par un trio d'agents de grands modèles de langage (LLM) complété par une baseline d'apprentissage automatique. Ce cadre répond aux quatre lacunes structurelles identifiées par Zafar et Wu dans leur revue systématique de 49 travaux portant sur la XAI en détection de fraude : prévalence des métriques de substitution prédictives, distorsions *data-centric* induites par le sur-échantillonnage synthétique, rareté de l'évaluation humaine, et monoculture méthodologique autour de SHAP.

Le cadre est instancié sur les jeux de données Elliptic Bitcoin et Ethereum Fraud, à travers cinq explicateurs (SHAP, LIME, Integrated Gradients, GNNExplainer, GraphLIME) et quatre classifieurs (Random Forest, LightGBM, Temporal-GCN, GraphSAGE). SHAP domine sur les modèles tabulaires, tandis qu'Integrated Gradients l'emporte sur les modèles à graphes. L'ajout d'explications aux invites soumises aux LLM fait passer l'exactitude décisionnelle de 0,52–0,79 à 0,81–0,94, et le coefficient kappa de Cohen entre agents de valeurs proches de zéro à des valeurs proches de un. Une fois les explications fournies, les décisions des trois agents LLM concordent presque parfaitement avec celles d'une forêt aléatoire entraînée sur les mêmes attributions, ce qui suggère, dans les limites de l'échantillon évalué (cinq instances par condition), un transfert du signal discriminant. Le mémoire étend l'article conférence par une revue de littérature élargie, une formalisation méthodologique approfondie et une discussion des implications opérationnelles.

Mots-clés : Intelligence artificielle explicable, blockchain, détection de fraude, réseaux de neurones sur graphes, grands modèles de langage, cadre d'évaluation, BRAS.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

AML	Anti-Money Laundering
AUC	Area Under the Curve
BRAS	Blockchain Rule Alignment Score
CoV	Coefficient of Variation
DA	Decision Accuracy
DVR	Dominant Violation Rate
ECE	Expected Calibration Error
EU	Explanation Utilisation
F1	F1-score (moyenne harmonique de la précision et du rappel)
FPR	False Positive Rate (taux de faux positifs)
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network (réseau de neurones sur graphe)
ICML	International Conference on Machine Learning
IEEE	Institute of Electrical and Electronics Engineers
IG	Integrated Gradients
IoT	Internet of Things (Internet des Objets)
KDD	Knowledge Discovery and Data Mining
LGB	LightGBM (Light Gradient Boosting Machine)
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model (grand modèle de langue)
MCC	Matthews Correlation Coefficient
ML	Machine Learning (apprentissage automatique)
MSE	Mean Squared Error (erreur quadratique moyenne)
NeurIPS	Neural Information Processing Systems
PR-AUC	Precision-Recall Area Under the Curve
RAS	Rule Alignment Score
RF	Random Forest
ROC	Receiver Operating Characteristic
RQ	Research Question (question de recherche)
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence (intelligence artificielle explicable)

DÉDICACE

À mes parents, dont la confiance silencieuse a guidé chacun de mes pas.

À ma famille et à mes proches, pour leur patience et leurs encouragements au long de ce parcours.

À celles et ceux qui croient encore que la rigueur scientifique est aussi une forme d'espérance.

REMERCIEMENTS

Je rends d'abord grâce à Dieu, l'Éternel Tout-Puissant, pour la force et la persévérance qu'il m'a accordées tout au long de ce travail. L'aboutissement d'un mémoire de maîtrise n'est jamais le fruit d'une démarche solitaire. Il s'inscrit dans une trajectoire collective faite de rencontres, de conseils et de soutiens qu'il convient de saluer avec gratitude.

J'adresse mes premiers remerciements à mon directeur de recherche, le professeur Fehmi Jaafar, dont l'encadrement scientifique a été à la fois exigeant et bienveillant. Sa lecture attentive de mes itérations successives, sa capacité à recadrer un raisonnement qui s'égare et sa confiance constante dans l'autonomie qu'il m'a accordée ont façonné, en profondeur, la manière dont je conçois aujourd'hui la recherche. Les nombreuses discussions tenues à l'UQAC, ainsi que les jalons posés lors de nos réunions périodiques, ont nourri à la fois la structure et l'ambition de ce travail.

Je remercie chaleureusement Wissam Salhab, collaborateur scientifique et co-auteur de l'article associé à ce mémoire, pour la qualité de ses retours méthodologiques. Ses observations sur la formalisation correcte de la constante de Lipschitz, ainsi que sur la simplification pédagogique des supports visuels, ont eu un impact direct et mesurable sur le manuscrit final. Travailler à ses côtés a été une chance, tant sur le plan scientifique que sur le plan humain.

Je tiens à remercier l'Université du Québec à Chicoutimi (UQAC) ainsi que le Département d'informatique et de mathématique, pour les ressources matérielles, l'environnement de recherche stimulant et l'ouverture aux problématiques industrielles qu'ils ont mis à ma disposition. Les échanges avec les enseignants chercheurs et les étudiants des cycles supérieurs, lors des séminaires et des groupes de lecture, ont régulièrement éclairé des points qui restaient pour moi en suspens.

Je salue également les collègues étudiants et chercheurs avec lesquels j'ai partagé bureaux, cafés et préoccupations académiques. Leurs questions, leurs critiques amicales et leur disponibilité ont contribué à clarifier mes idées et à maintenir, au long des mois de rédaction, une dynamique de travail saine.

Sur un plan plus personnel, je remercie ma famille, dont le soutien indéfectible, malgré la distance, a constitué la base affective sur laquelle s'est construit ce projet. Mes pensées vont en particulier à mes parents, dont l'éducation patiente m'a transmis le goût de la persévérance et de la curiosité intellectuelle. Je remercie enfin mes amis proches, qui ont su accueillir mes absences, mes doutes et mes enthousiasmes avec une bienveillance que je n'oublie pas.

Que toutes les personnes qui, de près ou de loin, ont contribué à la maturation de ce mémoire trouvent ici l'expression sincère de ma reconnaissance.

AVANT-PROPOS

Ce mémoire de maîtrise en informatique trouve son origine dans une interrogation simple, posée dès les premiers mois de recherche : comment juger sérieusement qu'une explication d'un modèle d'apprentissage est de bonne qualité, lorsque l'objet à expliquer est une décision automatique de détection de fraude sur une blockchain publique ? Cette question, en apparence modeste, ouvre un champ d'analyse à la croisée de plusieurs disciplines, l'apprentissage automatique, la cryptomonnaie, la conformité réglementaire et l'évaluation empirique des systèmes intelligents, et elle s'est progressivement structurée en un programme de recherche que ce manuscrit s'efforce de restituer.

Le socle scientifique du mémoire est constitué par un article intitulé *Beyond Predictive Metrics: A Multi-Dimensional Framework for XAI Evaluation in Blockchain Fraud Detection*, co-rédigé avec Wissam Salhab et le professeur Fehmi Jaafar, et accepté à la conférence *IEEE International Conference on Software Quality, Reliability, and Security* (QRS 2026), où il a été présenté. L'article propose un cadre d'évaluation à quatre modules, instancie ce cadre sur les jeux de données Elliptic Bitcoin et Ethereum Fraud, et rapporte un résultat empirique central : une fois munis d'une explication fidèle, des agents de grands modèles de langage prennent des décisions en accord presque parfait avec une forêt aléatoire entraînée sur les mêmes attributions. La contrainte de format imposée par la conférence (huit pages, double colonne) a toutefois laissé peu de place au déploiement complet du raisonnement, à la mise en perspective historique des métriques d'explicabilité, ou encore à la discussion détaillée des choix méthodologiques.

Le présent mémoire prolonge et étend cet article selon trois axes. Premièrement, il propose une revue de littérature substantiellement élargie, qui couvre non seulement l'état de l'art en XAI pour la fraude blockchain, mais également les cadres généralistes d'évaluation (Quantus, OpenXAI), les approches *LLM-as-judge* émergentes, et la cartographie des lacunes méthodologiques identifiées par Zafar et Wu. Deuxièmement, il formalise plus rigoureusement le cadre méthodologique, en explicitant les conventions de notation, les choix d'hyperparamètres et les justifications théoriques de chaque métrique. Troisièmement, il développe les discussions sur l'orthogonalité empirique des axes, les implications opérationnelles pour les cellules de conformité, ainsi que les perspectives ouvertes vers l'inclusion future d'experts humains et l'extension à d'autres écosystèmes blockchain.

CHAPITRE I

INTRODUCTION

1.1 CONTEXTE GÉNÉRAL

Les blockchains publiques règlent aujourd'hui plusieurs milliers de milliards de dollars de valeur par année. Bitcoin, fondé sur le modèle des sorties non dépensées (*Unspent Transaction Outputs*, UTXO), et Ethereum, fondé sur un modèle à comptes et permettant l'exécution de contrats intelligents, structurent l'essentiel de cet écosystème. Le *Crypto Crime Report* de Chainalysis Chainalysis (2024) estime que plusieurs dizaines de milliards de dollars d'activité illicite ont transité par ces chaînes en 2024, avec trois familles dominantes : les transferts associés à des adresses sanctionnées (notamment depuis la désignation de Tornado Cash par l'OFAC en août 2022), les flux liés à des groupes étatiques comme Lazarus, et les paiements de *ransomware* dont le cumul a dépassé le milliard de dollars annuel en 2023. Le démantèlement de places de marché comme Hydra en 2022 a partiellement redirigé l'activité plutôt que de l'éliminer, ce qui illustre la résilience structurelle du paysage criminel.

Cette demande s'inscrit dans le cadre élargi de la lutte contre le blanchiment d'argent et le financement du terrorisme (AML/CFT). Les recommandations du Groupe d'action financière (GAFI) Financial Action Task Force (2021) imposent depuis 2019 aux fournisseurs de services d'actifs virtuels des obligations de connaissance du client (KYC), de surveillance continue et de signalement, complétées en Europe par le règlement MiCA European Parliament and Council (2023) et aux États-Unis par les exigences du FinCEN. Dans cette architecture, une explication intelligible des décisions automatiques devient une condition de conformité.

Trois propriétés structurelles distinguent la fraude blockchain de la fraude bancaire classique. Le pseudonymat dissocie l'adresse de l'identité civile et n'admet qu'une attribution probabiliste par *clustering*, depuis Meiklejohn Meiklejohn *et al.* (2013) jusqu'à Möser et Narayanan Möser & Narayanan (2022). L'irrévocabilité interdit tout *chargeback* et impose une détection préventive. La vélocité permet la fragmentation et l'exfiltration en quelques minutes vers des juridictions non coopératives. À cela s'ajoute la composabilité des contrats intelligents qui ouvre la voie à des attaques sophistiquées (*flash loans*, MEV) Daian *et al.* (2020). La communauté a répondu par des classifieurs supervisés tabulaires (Random Forest Breiman (2001), LightGBM Ke *et al.* (2017)) et graphiques (GCN Kipf & Welling (2017), GraphSAGE Hamilton *et al.* (2017)), évalués sur les corpus Elliptic Bitcoin Weber *et al.* (2019) et Ethereum Fraud Farrugia *et al.* (2020).

1.2 PROBLÉMATIQUE SCIENTIFIQUE

La performance prédictive (F1, AUC) ne suffit pas à fonder une décision susceptible de geler des fonds ou de déclencher une enquête. Un analyste en conformité requiert une explication simultanément fidèle au modèle, stable sous perturbation, alignée avec le savoir métier et utile en aval. Or la revue systématique de Zafar et Wu Zafar & Wu (2026) sur 49 études XAI consacrées à la détection de fraude met en lumière quatre lacunes structurelles : près de 80 % des travaux évaluent la qualité explicative par le F1 du classifieur sous-jacent, le sur-échantillonnage SMOTE Chawla *et al.* (2002) dégrade la fidélité des explications, moins de 5 % impliquent un expert humain, et SHAP Lundberg & Lee (2017) concentre à lui seul 26 occurrences sur 49 au détriment de LIME Ribeiro *et al.* (2016), Integrated Gradients Sundararajan *et al.* (2017), GNNExplainer Ying *et al.* (2019) ou GraphLIME Huang *et al.* (2022).

Cette inadéquation est amplifiée par un corpus réglementaire (GAFI, MiCA, FinCEN) qui impose explicitement l'auditabilité algorithmique, et par une tension structurelle, théorisée par Rudin Rudin (2019) et observée empiriquement par Bhatt et collaborateurs Bhatt *et al.* (2020), entre performance des classifieurs profonds et transparence des modèles intrinsèquement interprétables. À ces difficultés s'ajoute une exigence propre au domaine : les explications doivent être interprétables au regard de la grammaire transactionnelle blockchain, ce qui implique l'incorporation de règles métier (concentration vers un mélangeur connu, sortie en poussière, périodicité suspecte). Aucun cadre académique existant ne propose, à notre connaissance, d'intégration formelle de cette connaissance domaine dans la mesure de la qualité explicative, alors que les plateformes commerciales (Chainalysis Reactor, Elliptic Investigator, TRM Labs) en font un usage central. La métrique BRAS proposée dans ce mémoire entend établir ce pont.

1.3 QUESTIONS DE RECHERCHE

Trois questions de recherche structurent le travail.

— **RQ1** : Comment évaluer la qualité des explications XAI au-delà de la performance prédictive du classifieur, en intégrant simultanément fidélité, stabilité et robustesse ?

— **RQ2** : Comment intégrer la connaissance du domaine blockchain (règles AML, indicateurs structurels, contre-indications expertes) dans une métrique opérationnelle d'évaluation des explications ?

— **RQ3** : Comment substituer, de manière scientifiquement défendable, à une étude utilisateur humaine coûteuse, un protocole automatisé fondé sur des agents de grands modèles de langage ?

Ces trois questions sont complémentaires plutôt que hiérarchiques : la première relève de l'évaluation algorithmique, la deuxième de l'évaluation *data-centric* et métier, la troisième de l'évaluation *human-centric* et de son automatisation.

1.4 HYPOTHÈSES

Chaque hypothèse est associée à une procédure expérimentale dont l'issue peut conduire à son rejet. Les tests statistiques applicables et conditions de rejet sont précisés au chapitre 4, les résultats au chapitre 6.

— **H1 (orthogonalité des axes)** : Les quatre axes proposés (fidélité, stabilité, alignement domaine, utilité en aval) capturent des dimensions empiriquement non redondantes de la qualité explicative. Concrètement, une méthode XAI pourra dominer sur un axe tout en étant dominée sur un autre, ce qui justifie l'évaluation conjointe.

— **H2 (paradoxe SMOTE-fidélité)** : Le sur-échantillonnage synthétique de type SMOTE préserve la performance prédictive mais dégrade significativement la fidélité des explications, alors que l'apprentissage *cost-sensitive* préserve à la fois la performance et la fidélité. Cette hypothèse opérationnalise le *paradoxe Zafar* dans le contexte blockchain.

— **H3 (effet des explications sur les LLM)** : L'ajout progressif d'information explicative dans les invites soumises à un agent LLM améliore la convergence inter-agents, mesurée par le coefficient kappa de Cohen, selon l'ordre $\kappa_{C3} > \kappa_{C2} > \kappa_{C1}$ où C1 fournit la prédiction seule, C2 ajoute l'explication, et C3 ajoute en plus une bande d'incertitude calibrée.

1.5 OBJECTIFS DU MÉMOIRE

OBJECTIF GÉNÉRAL

L'objectif général est de concevoir, formaliser, instancier et valider empiriquement un cadre multi-dimensionnel d'évaluation des méthodes XAI appliquées à la détection de fraude sur blockchains publiques, reproductible, applicable à des modèles tabulaires comme graphiques, et capable d'intégrer la connaissance experte et une dimension d'utilité *human-centric*.

OBJECTIFS SPÉCIFIQUES

1. **Formaliser** un cadre à quatre modules (Fidélité, Stabilité, BRAS, Utilité) couvrant les dimensions technique, métier et humaine, avec conventions de notation, normalisation et agrégation explicitées au chapitre 4.
2. **Concevoir et implémenter** la métrique originale BRAS (*Blockchain Rule Alignment Score*), combinant score d'alignement aux règles métier et taux de violation par contre-indications, instanciée sur deux catalogues distincts (Bitcoin Elliptic, Ethereum Fraud).
3. **Définir** un protocole d'évaluation de l'utilité fondé sur trois agents LLM hétérogènes (Claude, Gemini, GPT) interrogés sous trois conditions d'information (C1 prédiction seule, C2 + explication, C3 + incertitude calibrée), accompagné d'une baseline ML calibrée.
4. **Valider empiriquement** les trois hypothèses H1 à H3 par une étude comparative sur Elliptic Bitcoin et Ethereum Fraud, couvrant cinq méthodes XAI (SHAP, LIME, IG, GNNExplainer, GraphLIME) et quatre classifieurs (RF, LightGBM, T-GCN, GraphSAGE).

1.6 CONTRIBUTIONS DU MÉMOIRE

Le mémoire apporte quatre contributions principales. Premièrement, un *framework* d'évaluation à quatre modules qui unifie fidélité, stabilité, alignement domaine et utilité en aval dans un même pipeline reproductible, sans équivalent à notre connaissance dans le contexte blockchain. Deuxièmement, la métrique BRAS, qui opérationnalise la connaissance du domaine blockchain sous la forme d'un score d'alignement aux règles métier, complété par un taux de violation des contre-indications expertes. Troisièmement, un protocole *LLM-as-judge* structuré en trois conditions d'information, accompagné d'une baseline d'apprentissage automatique qui mesure dans quelle proportion l'explication transfère le signal discriminant du modèle enseignant vers le modèle évaluateur. Quatrièmement, une étude empirique systématique sur deux corpus blockchain hétérogènes, dont les résultats principaux ont été acceptés à QRS 2026 (où l'article a été présenté) et sont ici étendus et discutés en profondeur.

RQ1 trouve sa réponse principale dans les modules Fidélité et Stabilité, RQ2 dans la métrique BRAS, RQ3 dans le protocole LLM-as-judge avec baseline ML. Cette articulation conditionne l'organisation du chapitre 4 et des résultats au chapitre 6.

1.7 MÉTHODOLOGIE GÉNÉRALE

L'architecture méthodologique repose sur un pipeline en trois étapes. Premièrement, plusieurs classifieurs sont entraînés sur les deux jeux en privilégiant l'apprentissage *cost-sensitive* Elkan (2001) : Random Forest et LightGBM pour la modalité tabulaire, Temporal-GCN et GraphSAGE pour la modalité graphique disponible sur Elliptic. Le découpage est temporel (70/15/15 sur les 49 *timesteps* d'Elliptic) ou stratifié (Ethereum), avec recherche d'hyperparamètres par validation croisée temporelle et graines aléatoires

fixées. Deuxièmement, chaque modèle est interrogé par les explicateurs compatibles avec sa nature (SHAP et LIME pour le tabulaire ; IG, GNNExplainer et GraphLIME pour le graphique). Troisièmement, le cadre à quatre modules est appliqué systématiquement aux paires (modèle, explicateur) : le module Fidélité agrège *comprehensiveness*, *sufficiency* et infidelity de Yeh Yeh *et al.* (2019) ; le module Stabilité combine Lipschitz local Alvarez-Melis & Jaakkola (2018), corrélation de Kendall, coefficient de variation par bootstrap et test d'identité ; le module BRAS confronte la liste ordonnée des variables importantes à un catalogue de règles propre à chaque corpus ; le module Utilité confronte la décision des agents LLM, sous C1-C3, à une baseline ML calibrée, mesurée par le kappa de Cohen Cohen (1960). Les sorties normalisées dans $[0, 1]$ et agrégées par moyenne pondérée produisent un vecteur quadridimensionnel comparable.

1.8 ORGANISATION DU MÉMOIRE

Le chapitre 2 fournit l'arrière-plan technique (blockchain, classifieurs tabulaires et graphiques, taxonomie XAI, métriques, LLM). Le chapitre 3 organise l'état de l'art selon cinq axes et explicite le positionnement face aux lacunes de Zafar et Wu. Le chapitre 4 formalise le cadre à quatre modules. Le chapitre 5 décrit le dispositif expérimental (jeux de données, hyperparamètres, découpage, graines). Le chapitre 6 présente les résultats module par module et le classement comparatif. Le chapitre 7 examine les implications théoriques et opérationnelles puis expose les menaces à la validité (section 7.6). Le chapitre 8 synthétise les apports et l'agenda doctoral. Plusieurs annexes détaillent les catalogues BRAS, les invites LLM, les configurations et le dépôt de code reproductible.

CHAPITRE II

CONCEPTS FONDAMENTAUX

Ce chapitre rassemble les notions techniques nécessaires aux chapitres ultérieurs, en suivant une progression du concret (blockchain, fraude) vers l'abstrait (architectures de classification, explicabilité, métriques d'évaluation, LLM évaluateurs). Cette progression reflète la chaîne de dépendances conceptuelles : un explicateur s'applique à un classifieur, une métrique s'applique à un explicateur, et un LLM agit en aval de l'explication.

2.1 BLOCKCHAIN ET ÉCOSYSTÈME DE LA FRAUDE

2.1.1 ARCHITECTURES UTXO ET COMPTE

Une blockchain publique est un grand livre distribué où les transactions sont regroupées en blocs chaînés par des empreintes cryptographiques. Deux familles d'architectures dominent. Le modèle UTXO, popularisé par Bitcoin, conçoit une transaction comme la consommation de sorties non dépensées et la production de nouvelles sorties ; une adresse n'a pas de solde explicite et se représente naturellement en graphe biparti (transactions, adresses). Le modèle à comptes, adopté par Ethereum, traite chaque adresse comme un compte porteur d'un solde, mis à jour atomiquement, et facilite l'exécution de contrats intelligents avec une représentation en graphe orienté entre comptes. Sur le plan de l'extraction de caractéristiques, le modèle UTXO favorise les attributs liés au sous-graphe transactionnel (degrés, profondeurs, motifs), tandis que le modèle à comptes encourage les attributs agrégés par compte (nombre de transactions, valeur cumulée, périodicité).

2.1.2 TYPOLOGIES DE FRAUDE ET TECHNIQUES D'OBFUSCATION

Les fraudes blockchain se classent en plusieurs typologies qui se recouvrent partiellement : blanchiment AML par fragmentation et recomposition des flux, hameçonnage (*phishing*) par usurpation des clés, schémas de Ponzi dans la finance décentralisée, *rug pulls* (retrait brutal des fonds après afflux initial), exploitation de contrats vulnérables, et MEV (*sandwich attacks*, extraction de valeur par les producteurs de blocs) Daian *et al.* (2020). Les services de mélange brouillent l'origine des fonds : première génération par agrégation volontaire (CoinJoin, Wasabi, Samourai), deuxième génération par preuves à divulgation nulle (Tornado Cash sur Ethereum, dont l'inscription sur la liste OFAC en août 2022 a fait du code lui-même l'objet de la sanction). Möser et Narayanan Möser & Narayanan (2022) montrent qu'une fraction substantielle des flux mélangés reste partiellement attribuable par *clustering* amélioré.

2.1.3 PSEUDONYMAT ET DÉSANONYMISATION

Le pseudonymat n'est pas l'anonymat absolu. Depuis Meiklejohn Meiklejohn *et al.* (2013), la littérature mobilise principalement deux heuristiques de *clustering* : l'entrée commune (les adresses fournissant les entrées d'une transaction sont contrôlées par la même entité) et l'adresse de changement (identification de la sortie retournant à l'expéditeur). Ces heuristiques, complétées par l'étiquetage de services connus, alimentent les bases commerciales (Chainalysis, Elliptic, TRM Labs) dont les sorties servent de référence au jeu Elliptic utilisé dans ce mémoire.

2.1.4 CADRE AML/CFT

Le cadre AML/CFT international repose sur les recommandations du GAFI, dont la *Travel Rule* (recommandation 16) étendue en 2019 aux fournisseurs de services d'actifs virtuels impose la transmission d'informations sur l'identité des parties pour toute transaction au-dessus d'un seuil. Les obligations KYC, de surveillance continue et de signalement aux cellules de renseignement financier (CANAFE au Canada, transposition AMLD en Europe) structurent l'action des plateformes. Cet enjeu réglementaire motive à la fois la production d'explications intelligibles et la définition du catalogue de règles métier mobilisé par BRAS.

2.2 APPRENTISSAGE SUPERVISÉ POUR LA DÉTECTION

2.2.1 RANDOM FOREST

Les forêts aléatoires de Breiman Breiman (2001) combinent un grand nombre d'arbres entraînés sur des sous-échantillons *bootstrap*, avec restriction aléatoire des variables candidates à chaque nœud (typiquement $\sqrt{p}$ sur p variables disponibles). L'agrégation par vote majoritaire réduit la variance des arbres tout en préservant leur faible biais. Les instances non utilisées (*out-of-bag*, environ 37 % en moyenne) permettent une estimation de l'erreur sans validation croisée externe. Cette robustesse, conjuguée au traitement direct d'attributs hétérogènes, explique la popularité de RF en détection tabulaire et autorise une explication post-hoc efficace via TreeSHAP.

2.2.2 GRADIENT BOOSTING (LIGHTGBM)

Le *gradient boosting* construit séquentiellement un ensemble d'arbres faibles, chacun corrigeant les résidus du précédent dans la direction du gradient. LightGBM Ke *et al.* (2017)

se distingue par la croissance par feuille (*leaf-wise*, qui privilégie la feuille au plus fort gain et requiert le contrôle explicite de `num_leaves`), le *binning* par histogrammes (discrétisation en environ 255 seaux pour une évaluation linéaire des coupures), l'échantillonnage stratégique d'instances (*Gradient-based One-Side Sampling*) et le regroupement de variables mutuellement exclusives. Ces choix rendent LightGBM particulièrement efficace sur de grands jeux tabulaires déséquilibrés.

2.2.3 SHAP POUR MODÈLES À BASE D'ARBRES

La valeur de Shapley d'une variable i dans une fonction de gain v s'écrit

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

dont l'évaluation directe est exponentielle. TreeSHAP Lundberg & Lee (2017) exploite la structure récursive des ensembles d'arbres pour calculer ces valeurs en temps polynomial, ce qui le rend attractif pour RF et LightGBM. KernelSHAP en fournit une variante agnostique applicable à tout modèle, au prix d'une variance d'estimation plus élevée. Dans ce mémoire, TreeSHAP est employé pour les modèles à arbres et KernelSHAP n'est invoqué qu'à titre comparatif dans le module Stabilité.

2.2.4 APPRENTISSAGE COST-SENSITIVE

Le déséquilibre de classes, structurellement présent en détection de fraude, peut être traité soit par l'apprentissage *cost-sensitive* Elkan (2001) (affectation d'un poids plus élevé aux erreurs sur la classe minoritaire au niveau de la perte), soit par le rééquilibrage SMOTE Chawla *et al.* (2002) (génération synthétique par interpolation entre instances minoritaires voisines). Si les deux conduisent à des F1 comparables, la revue de Zafar et

Wu Zafar & Wu (2026) suggère que SMOTE détériore la fidélité des explications post-hoc, ce qui justifie l'adoption systématique du *cost-sensitive* dans le présent travail.

2.3 GRAPH NEURAL NETWORKS

2.3.1 GCN, GRAPHSAGE ET CADRE DE MESSAGE PASSING

Les GNN généralisent la convolution au cas d'un graphe muni d'attributs sur ses nœuds. Les *Graph Convolutional Networks* Kipf & Welling (2017) reposent sur une approximation au premier ordre de la convolution spectrale : chaque représentation est mise à jour par combinaison linéaire des voisines suivie d'une non-linéarité, ce qui est intrinsèquement transductif. GraphSAGE Hamilton *et al.* (2017) introduit une agrégation inductive : chaque nœud échantillonne un sous-ensemble fixe de voisins et combine leurs représentations (moyenne, LSTM, max-pooling), ce qui permet de classer des nœuds inconnus à l'entraînement, scénario précisément correspondant à la détection en flux.

Le cadre *Message Passing Neural Network* Gilmer *et al.* (2017) unifie ces architectures : pour chaque nœud v à la couche t, une phase de message produit $m_{u\to v}^{(t)} = M_t(h_v^{(t-1)}, h_u^{(t-1)}, e_{u,v})$, et une phase d'agrégation et de mise à jour combine ces messages, $h_v^{(t)} = U_t(h_v^{(t-1)}, \mathrm{AGG}(\{m_{u\to v}^{(t)}\}))$. Ce formalisme absorbe les GCN, GraphSAGE et les GAT Veličković *et al.* (2018) (qui introduisent une attention apprise pondérant dynamiquement les voisins), et fournit le socle d'explicateurs comme GNNExplainer dont le masque opère sur les messages. Sur le plan de la passage à l'échelle, GraphSAGE limite par échantillonnage le nombre de voisins par couche, ce qui borne la complexité par mini-batch et autorise l'entraînement sur des graphes de plusieurs millions de nœuds.

2.3.2 TEMPORAL-GCN

La structure temporelle des transactions porte de l'information : une même adresse peut être légitime sur une période et frauduleuse sur une autre. T-GCN combine un GCN spatial et une GRU temporelle, traitant chaque *timestep* comme un graphe à part entière ; d'autres variantes (EvolveGCN, DySAT) font évoluer les paramètres ou ajoutent une attention spatio-temporelle. Pour Elliptic (49 *timesteps*), le choix se porte sur T-GCN en raison de sa simplicité, de sa reproductibilité et d'une implémentation de référence disponible dans PyTorch Geometric. Le découpage adopté est temporel (70/15/15) Weber *et al.* (2019). Pour Ethereum, faute de structure temporelle exploitable au niveau du compte, un découpage stratifié est employé.

2.4 EXPLICABILITÉ (XAI)

2.4.1 TAXONOMIE

La XAI s'organise selon plusieurs axes orthogonaux Doshi-Velez & Kim (2017); Arrieta *et al.* (2020); Guidotti *et al.* (2018) : *post-hoc* versus intrinsèque (explication a posteriori contre modèles intrinsèquement interprétables comme régression linéaire, arbre court ou modèles additifs généralisés) ; *local* versus global (instance particulière contre comportement d'ensemble) ; *spécifique au modèle* versus agnostique (TreeSHAP, Grad-CAM contre KernelSHAP, LIME). À ces dimensions s'ajoutent les distinctions secondaires entre explication par variables, par exemples, par contre-factuels ou par règles symboliques. La tension fondamentale entre fidélité au modèle et intelligibilité pour l'utilisateur structure le champ Lipton (2018); Rudin (2019). Les méthodes retenues ici (SHAP, LIME, IG, GNNExplainer, GraphLIME) appartiennent au quadrant post-hoc, local, par attribution, mais se répartissent entre spécifique et agnostique.

2.4.2 SHAP

SHAP Lundberg & Lee (2017) repose sur les valeurs de Shapley issues de la théorie des jeux coopératifs et fournit une décomposition additive de la prédiction satisfaisant les axiomes d'efficacité, de symétrie, de *dummy* et d'additivité. TreeSHAP en donne une variante exacte et polynomiale pour les modèles à arbres, KernelSHAP une variante agnostique plus coûteuse.

2.4.3 LIME

LIME Ribeiro *et al.* (2016) construit un modèle interprétable local autour d'une instance à expliquer. Un voisinage est échantillonné par perturbation, le modèle initial est évalué dessus, puis un modèle linéaire pondéré est ajusté. Formellement, l'explication s'obtient par

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

où $\pi_x(z) = \exp(-d(x,z)^2/\sigma^2)$ pondère les échantillons et $\Omega(g)$ pénalise la complexité. LIME présente l'avantage de l'agnosticité totale mais souffre d'une stabilité notoirement faible, en raison de l'aléa d'échantillonnage Slack *et al.* (2020); Adebayo *et al.* (2018).

2.4.4 INTEGRATED GRADIENTS

IG Sundararajan *et al.* (2017) attribue à chaque variable une contribution proportionnelle à l'intégrale du gradient le long d'un chemin reliant une référence (*baseline*) à l'instance. Cette méthode, spécifique aux modèles différentiables, satisfait par construction deux axiomes fondamentaux : la complétude (la somme des attributions reproduit l'écart de prédiction entre instance et référence) et la sensibilité (toute variable causalement importante reçoit une attribution non nulle). Elle satisfait également l'*Implementation Invariance*

(deux modèles fonctionnellement équivalents reçoivent les mêmes attributions), ce qui en fait le candidat privilégié pour les modèles différentiables.

2.4.5 EXPLICATEURS POUR GRAPHES

GNNExplainer Ying *et al.* (2019) optimise conjointement un masque continu sur les arêtes et un masque sur les variables nodales pour maximiser l'information mutuelle $I(Y;(G_S,X_S))$ entre la prédiction conservée et le sous-graphe explicatif induit. GraphLIME Huang *et al.* (2022) étend LIME aux graphes par une régression HSIC-Lasso non linéaire. PGExplainer Luo *et al.* (2020) et SubgraphX Yuan *et al.* (2021) enrichissent le paysage, comme le résume Yuan et collaborateurs Yuan *et al.* (2022).

2.4.6 ATTAQUES ADVERSARIALES SUR LES EXPLICATEURS

Slack et collaborateurs Slack *et al.* (2020) montrent par *scaffolding* qu'un classifieur biaisé peut être construit pour présenter aux explicateurs un comportement apparent débarrassé du biais : en détectant si l'instance provient des perturbations utilisées par SHAP ou LIME, le modèle bascule dans un comportement innocent dans ce cas. Adebayo et collaborateurs Adebayo *et al.* (2018) montrent par ailleurs que plusieurs cartes de saillance restent inchangées sous randomisation complète du modèle. Ces fragilités motivent l'inclusion systématique d'une dimension de stabilité.

2.5 MÉTRIQUES D'ÉVALUATION DES EXPLICATIONS

2.5.1 FIDÉLITÉ

La fidélité quantifie le degré auquel l'explication reflète le comportement du modèle. Soit f le modèle, x une instance, $\phi(x) \in \mathbb{R}^d$ les attributions, et $\mathcal{I}_k$ l'ensemble des k variables

maximales. La *comprehensiveness*

$$\mathrm{Comp}_k(x) = f(x)_y - f(x_{\setminus \mathscr{I}_k})_y$$

mesure la chute de probabilité prédite lorsque ces variables sont masquées ; la *sufficiency*

$$\mathrm{Suff}_k(x) = f(x)_y - f(x_{\mathscr{I}_k})_y$$

mesure la préservation de la prédiction lorsque seules ces variables sont conservées. L'*Infidelity* Yeh *et al.* (2019) mesure $\mathbb{E}_\delta\left[(\phi(x)^\top \delta - (f(x) - f(x-\delta)))^2\right]$, soit l'écart quadratique entre projection de l'attribution sur une perturbation et variation effective de la prédiction. Ces trois mesures complémentaires fondent l'axe Fidélité.

2.5.2 STABILITÉ

La stabilité combine quatre indicateurs Alvarez-Melis & Jaakkola (2018); Adebayo *et al.* (2018) : la constante de Lipschitz locale $L(x) = \max_{x' \in \mathscr{B}_\varepsilon(x)} \|\phi(x') - \phi(x)\| / \|x' - x\|$ sur un voisinage de rayon ε ; la corrélation de Kendall τ sur les rangs d'attributions entre instances voisines ; le coefficient de variation par bootstrap (particulièrement pertinent pour LIME et GraphLIME) ; et un test d'identité élémentaire par ré-exécution stricte avec la même graine.

2.5.3 ALIGNEMENT DOMAINE ET AUTRES DIMENSIONS

Plusieurs travaux Zhou *et al.* (2021); Nauta *et al.* (2023) cartographient les dimensions plus qualitatives (plausibilité, compréhensibilité, alignement domaine) et soulignent leur sous-représentation. Ce mémoire opérationnalise spécifiquement l'alignement domaine

via la métrique BRAS au chapitre 4. La plausibilité et la compréhensibilité ne sont pas opérationnalisées comme modules autonomes, pour trois raisons : leur mesure suppose un panel humain qui dépasse le cadre du mémoire, leurs proxies automatisés (longueur, sparsité, lisibilité) capturent surtout du bruit, et l'alignement domaine via BRAS recouvre partiellement la plausibilité tandis que l'utilité approximée par LLM capture une fraction de la compréhensibilité.

2.6 LARGE LANGUAGE MODELS COMME ÉVALUATEURS

2.6.1 CAPACITÉS DE RAISONNEMENT ET CALIBRATION

Depuis GPT-3 Brown *et al.* (2020), les LLM ont démontré un raisonnement *zero-shot* sur des tâches structurées, amplifié par les invites *chain-of-thought* Kojima *et al.* (2022). La littérature distingue trois régimes : zero-shot pur, few-shot (quelques exemples), chain-of-thought (raisonnement intermédiaire explicite). Le présent mémoire adopte un *few-shot* contraint à sortie structurée, arbitrage entre qualité du raisonnement et reproductibilité du verdict (invites en annexe).

La calibration mesure l'écart entre confiance subjective et exactitude observée. L'Erreur de Calibration Attendue (ECE) Guo *et al.* (2017) discrétise les probabilités en M groupes (typiquement $M = 10$ ou 15) et somme l'écart pondéré entre confiance moyenne et exactitude. Pour les LLM, Kadavath et collaborateurs Kadavath *et al.* (2022) ont montré une calibration imparfaite mais informative. Pour les classifieurs du module Utilité, une calibration par régression isotonique sur la partition de validation rend les probabilités directement comparables aux estimations subjectives des agents, et garantit que la bande d'incertitude C3 a une interprétation probabiliste rigoureuse.

2.6.2 LLM-AS-JUDGE ET SES LIMITES

L'usage des LLM comme évaluateurs (*LLM-as-judge*) s'est généralisé. Zheng et collaborateurs Zheng *et al.* (2023) proposent MT-Bench (questions multi-tours évaluées par un LLM tiers) et Chatbot Arena (préférences humaines en double aveugle pour calibrage), et rapportent une forte corrélation entre jugements GPT-4 et préférences humaines agrégées. Liu et collaborateurs Liu *et al.* (2023) proposent G-Eval, fondé sur une chaîne de pensée explicite et une normalisation des scores, qui surpasse les métriques automatiques classiques (BLEU, ROUGE, BERTScore).

Plusieurs biais documentés affectent la pratique du LLM-as-judge (biais de position, biais de longueur, hallucinations, consensus artificiel entre modèles d'une même famille). Le protocole proposé atténue ces risques par la diversification des fournisseurs (Claude d'Anthropic, Gemini de Google, GPT d'OpenAI), la randomisation de l'ordre des conditions, la normalisation des invites et l'introduction d'une baseline ML qui ancre les conclusions sur un référent non linguistique reproductible.

CHAPITRE III

REVUE DE LITTÉRATURE

Ce chapitre dresse l'état de l'art selon cinq axes : fraude bancaire (terrain de maturation initial), fraude blockchain, cadres généralistes d'évaluation, approches *LLM-as-judge*, et synthèse des lacunes systémiques identifiées par Zafar et Wu Zafar & Wu (2026). Il se conclut par le positionnement du présent travail.

3.1 XAI EN DÉTECTION DE FRAUDE BANCAIRE

La détection de fraude bancaire a constitué le terrain initial d'application de la XAI. Psychoula et collaborateurs Psychoula *et al.* (2021) proposent une étude systématique de SHAP et LIME sur la fraude par carte bancaire, mettant déjà en évidence la fragilité des explications face aux choix de prétraitement. Liu et collaborateurs Liu *et al.* (2021) explorent la détection graphique par méta-graphes explicatifs, identifiant des motifs structurels (cycles courts, étoiles convergentes) intelligibles à un analyste. Les systèmes xFraud Rao *et al.* (2021) et SEFraud Li *et al.* (2024) intègrent l'explication au cœur du détecteur via un sous-graphe de raisonnement ou un masque appris, échappant à la critique de la fidélité post-hoc mais imposant une architecture spécifique. Demertzis et collaborateurs Demertzis *et al.* (2021) soulignent les exigences forensiques propres à la cybersécurité (traçabilité, défense contradictoire). Bhatt et collaborateurs Bhatt *et al.* (2020) relèvent par entretiens un décalage récurrent entre ce que les data scientists produisent et ce que les parties prenantes attendent. Slack et collaborateurs Slack *et al.* (2020) montrent que SHAP et LIME peuvent être adversariellement trompés ; Rudin Rudin (2019) défend l'abandon des post-hoc au profit des modèles intrinsèquement interprétables ; Adebayo et collaborateurs Adebayo *et al.* (2018) proposent des tests de cohérence préalables.

TABLEAU 3.1 : Travaux représentatifs en XAI pour la fraude bancaire.

Travail	Dataset	Modèle	XAI évalué
Psychoula et al. (2021)	Carte bancaire	RF, GBDT	SHAP, LIME
Rao et al. (2021, xFraud)	E-commerce industriel	GNN	Self-explainable
Li et al. (2024, SEFraud)	Multi-corpus	GNN	Masque appris
Bhatt et al. (2020)	Enquête praticiens	N/A	Synthèse qualitative
Slack et al. (2020)	Compas, Adult	RF, XGBoost	SHAP, LIME (adversarial)

Plusieurs limites se dégagent : la validation utilisateur reste exceptionnelle, la surréliance sur SHAP appauvrit les comparaisons, et les métriques rapportées restent dominées par F1 et AUC du classifieur plutôt que par une évaluation de l'explication. La littérature bancaire converge sur un constat : la XAI est devenue une exigence opérationnelle, mais son évaluation reste peu standardisée.

3.2 XAI EN DÉTECTION DE FRAUDE BLOCKCHAIN

3.2.1 DATASETS BLOCKCHAIN DE RÉFÉRENCE

Le tableau 3.2 récapitule les principaux corpus publics. Elliptic Bitcoin reste le plus utilisé pour les approches graphiques grâce à sa structure temporelle explicite et à un étiquetage manuel de qualité. Ethereum Fraud constitue le pendant tabulaire incontournable au niveau du compte, avec un déséquilibre prononcé pertinent pour les techniques cost-sensitive. Les datasets IBM Bitcoin AML, synthétiques, fournissent une vérité terrain contrôlée utile au développement mais limitée par un biais de réalisme.

3.2.2 APPROCHES GNN POUR L'AML BITCOIN

Weber et collaborateurs Weber *et al.* (2019) ont marqué un point d'inflexion en publiant Elliptic et un benchmark exploitant un GCN, observant un compromis entre performance et exploitation de l'information structurelle. Le jeu présente plusieurs spécificités

TABLEAU 3.2 : Datasets publics pour la détection de fraude blockchain.

Dataset	Chaîne	Taille	Structure	Étiquettes
Elliptic Bitcoin	Bitcoin	203 769 tx	Graphe + 49 timesteps	Illicite / licite / inconnu
Elliptic++	Bitcoin	Étendu	Multi-niveau (tx, adresse)	Multi-classes
Ethereum Fraud Farrugia *et al.* (2020)	Ethereum	9 841 comptes	Tabulaire	Frauduleux / normal
IBM Bitcoin AML	Synthétique	Configurable	Graphe	Catégories AML
Phishing Ethereum Chen *et al.* (2020)	Ethereum	∼2 000 comptes	Graphe + tabulaire	Phishing / normal

structurelles : 49 *timesteps* d'environ deux jours, fraction étiquetée de 23 % imposant un traitement semi-supervisé ou restreint aux étiquetées, ratio de classes d'environ 2 % d'illicites, et 166 caractéristiques (94 propres et 72 agrégées sur le voisinage). Alarab et collaborateurs Alarab *et al.* (2020) et Pocher et collaborateurs Pocher *et al.* (2023) prolongent cette ligne, sans toutefois évaluer la qualité explicative. Pareja et collaborateurs Pareja *et al.* (2020) proposent EvolveGCN qui apprend l'évolution temporelle des paramètres, améliorant marginalement le F1 illicite sur Elliptic mais sans évaluation explicative.

3.2.3 APPROCHES TABULAIRES POUR ETHEREUM

Farrugia et collaborateurs Farrugia *et al.* (2020) ouvrent la voie avec un jeu de 9 841 comptes et un benchmark XGBoost/LightGBM atteignant un F1 macro supérieur à 0,95. Chen et collaborateurs Chen *et al.* (2020) se spécialisent sur le phishing par combinaison de graphe transactionnel et de classifieurs supervisés. Goyal et collaborateurs Goyal *et al.* (2022) étendent le panel comparé et soulignent la sensibilité des résultats au choix de partition. L'évaluation explicative reste absente de ces travaux. Sur les explicateurs spécifiques aux graphes, GNNExplainer Ying *et al.* (2019) reste la référence, complétée

par PGExplainer Luo *et al.* (2020), SubgraphX Yuan *et al.* (2021) et GraphLIME Huang *et al.* (2022) Yuan *et al.* (2022).

3.2.4 LACUNES SPÉCIFIQUES AU DOMAINE BLOCKCHAIN

À notre connaissance, aucun travail antérieur ne propose une évaluation systématique de GNNExplainer, PGExplainer, SubgraphX et GraphLIME sur Elliptic selon un protocole reproductible : les rares comparaisons procèdent sur graphes synthétiques (BA-Shapes, Tree-Cycles), corpus moléculaires (MUTAG, BBBP) ou citations académiques (Cora, CiteSeer, PubMed), sémantiquement éloignés. Trois spécificités aggravent l'insuffisance des cadres existants : la nature hybride des corpus (graphique pour Bitcoin, tabulaire pour Ethereum) impose un cadre à double modalité, la connaissance métier (typologies AML, indicateurs structurels, contre-indications) est centrale et son intégration formelle absente, et la rareté des étiquettes limite la confiance par instance. Aucun travail à notre connaissance ne propose, sur les mêmes corpus, une évaluation conjointe de fidélité, stabilité, alignement domaine et utilité en aval.

3.3 FRAMEWORKS D'ÉVALUATION XAI

Quantus Hedström *et al.* (2023) fournit une bibliothèque ouverte regroupant plus de trente métriques organisées en six catégories : fidélité (Faithfulness Correlation, Pixel-Flipping, Region Perturbation), robustesse (Local Lipschitz Estimate, Avg-Sensitivity, Max-Sensitivity), localisation (Pointing Game, spécifique vision), complexité (sparsité, entropie), randomisation (dégradation sous randomisation, suite à Adebayo), et axiomatique (Completeness, Symmetry). L'outil impose un format PyTorch qui rend coûteuse son adaptation aux modèles à arbres et aux modèles graphiques.

OpenXAI Agarwal *et al.* (2022) organise un *leaderboard* public sur cinq dimensions (faithfulness, stability, fairness, sparsity, infidelity), sept jeux tabulaires (German Credit, Adult, COMPAS, etc.) et quatre classifieurs, révélant par exemple la supériorité systématique de KernelSHAP sur LIME en faithfulness. Ses limites sont l'absence de support graphique, la non-intégration d'une dimension domaine et le périmètre exclusivement tabulaire.

Captum Kokhlikyan *et al.* (2020) fournit une suite d'explicateurs pour PyTorch (IG, DeepLIFT, GradientShap, Occlusion, LRP) et des métriques rudimentaires ; Alibi Explain Klaise *et al.* (2021) se distingue par les explications par contre-factuels et prototypes. Aucun de ces outils ne couvre conjointement les quatre dimensions du cadre proposé.

TABLEAU 3.3 : Comparaison des principaux frameworks d'évaluation XAI.

Framework	Fidélité	Stabilité	Tabulaire	Graphe	Domaine
Quantus	Oui	Oui	Partiel	Partiel	Non
OpenXAI	Oui	Oui	Oui	Non	Non
Captum	Partiel	Partiel	Oui	Limité	Non
Alibi Explain	Non	Non	Oui	Non	Non
Notre cadre	Oui	Oui	Oui	Oui	BRAS

Sur le plan conceptuel, Doshi-Velez et Kim Doshi-Velez & Kim (2017) distinguent trois niveaux d'évaluation (centré application, humain, fonction), et Lipton Lipton (2018) décompose l'interprétabilité en simulabilité, décomposabilité et transparence algorithmique. La revue systématique de Nauta et collaborateurs Nauta *et al.* (2023) sur 312 articles révèle qu'un tiers seulement recourt à une métrique quantitative et que cinq métriques (Faithfulness, Sensitivity, AUC du masque, ROAR, Pointing Game) couvrent plus de 70 % des occurrences, suggérant un effet de dominance par disponibilité plutôt que par qualité intrinsèque Guidotti *et al.* (2018); Arrieta *et al.* (2020); Zhou *et al.* (2021). Le présent mémoire hérite de Quantus et OpenXAI sur les axes fidélité (Comprehensiveness, Sufficiency,

Infidelity de Yeh) et stabilité (Lipschitz, sensibilité moyenne) tout en introduisant deux modules originaux : l'alignement domaine via BRAS et l'utilité approximée par LLM.

3.4 APPROCHES LLM-AS-JUDGE

Les capacités *zero-shot* de Brown et collaborateurs Brown *et al.* (2020), amplifiées par Kojima et collaborateurs Kojima *et al.* (2022), fondent la pratique du LLM-as-judge. Zheng et collaborateurs Zheng *et al.* (2023) introduisent MT-Bench avec GPT-4 comme juge et rapportent une corrélation de plus de 80 % avec les préférences humaines de Chatbot Arena. Liu et collaborateurs Liu *et al.* (2023) proposent G-Eval avec chaîne de pensée structurée, qui surpasse BLEU, ROUGE et BERTScore en corrélation avec le jugement humain. Naiseh et collaborateurs Naiseh *et al.* (2023) étudient l'impact des explications sur la calibration de la confiance dans le domaine clinique. Kadavath et collaborateurs Kadavath *et al.* (2022) montrent par ailleurs que les LLM disposent d'une calibration imparfaite mais informative.

L'application à l'évaluation d'explications XAI reste peu consolidée. Trois axes structurent les protocoles : juge unique versus panel, avec versus sans référence, score scalaire versus décision discrète. Le présent mémoire se positionne dans le quadrant panel / sans référence / décision discrète, justifié par la nature opérationnelle de la décision (signaler ou non) et la volonté de mesurer l'accord inter-agents par kappa de Cohen. À notre connaissance, aucun travail n'a proposé d'utiliser des agents LLM comme proxies décisionnels en aval d'une explication XAI dans le contexte spécifique de la fraude blockchain, accompagnés d'une baseline ML servant de référent non linguistique reproductible. C'est l'apport du module Utilité du présent mémoire.

3.5 LIMITES OBSERVÉES ET GAPS DE LA LITTÉRATURE

La revue de Zafar et Wu Zafar & Wu (2026) sur 49 études XAI consacrées à la détection de fraude dégage quatre lacunes structurelles qui structurent le positionnement du mémoire.

3.5.1 GAP ÉVALUATION

Près de 80 % des études évaluent la qualité des explications par F1 ou AUC du classifieur sous-jacent, métriques qui ne portent pas sur l'explication mais sur le modèle expliqué. Cette confusion empêche toute méta-analyse rigoureuse et expose à un risque juridique non maîtrisé dans un contexte réglementaire où l'explication participe à la décision auditable. Le mémoire répond par un module Fidélité agrégé (comprehensiveness, sufficiency, infidelity) appliqué à toutes les paires (modèle, explicateur).

3.5.2 GAP DATA-CENTRIC

Plus de 60 % des études recourent à SMOTE sans s'interroger sur ses effets sur la qualité explicative. Le mécanisme est clair : en interpolant linéairement entre instances minoritaires voisines, SMOTE génère des points synthétiques pouvant occuper des régions où le modèle n'a pas appris de comportement cohérent, créant des frontières artefactuelles que les explicateurs post-hoc captent par perturbation. Le mémoire valide empiriquement ce *paradoxe SMOTE* sur Elliptic et Ethereum, et adopte systématiquement l'apprentissage cost-sensitive.

3.5.3 GAP HUMAN-CENTRIC

Moins de 5 % des études impliquent un expert humain. Sur les 2 ou 3 études qui mobilisent effectivement des évaluateurs, le panel moyen est inférieur à dix participants avec une qualification métier rarement documentée. Cette lacune prive la communauté de toute mesure d'utilité réelle et empêche la calibration des métriques automatiques sur un référent humain validé. Le mémoire propose un proxy automatisé fondé sur trois agents LLM hétérogènes, complété par une baseline ML, sans prétendre se substituer à une étude humaine ultérieure inscrite à l'agenda doctoral.

3.5.4 GAP MONOCULTURE

SHAP apparaît dans 26 des 49 études, soit plus de la moitié du corpus. Cette monoculture, alimentée par un effet de réseau (popularité initiale, outillage Python performant, coût d'usage relatif réduit), biaise les classements vers les configurations où SHAP performe bien, sous-évalue les explicateurs spécifiques aux graphes (GNNExplainer, GraphLIME, PGExplainer) sur des corpus réels et freine leur adoption opérationnelle. Le mémoire compare systématiquement cinq explicateurs sur deux datasets et quatre classifieurs.

3.6 POSITIONNEMENT DU PRÉSENT MÉMOIRE

Le positionnement se résume en quatre points. Premièrement, le cadre répond simultanément aux quatre lacunes de Zafar et Wu : fidélité (gap évaluation), comparaison cost-sensitive versus SMOTE (data-centric), proxy LLM (human-centric), cinq explicateurs hétérogènes (monoculture). Deuxièmement, la métrique BRAS formalise pour la première fois l'intégration de règles métier blockchain dans l'évaluation explicative. Troisièmement, le protocole LLM-as-judge est ancré sur une baseline ML, ce qui renforce sa validité et

mesure le transfert de signal entre modèle enseignant et évaluateur. Quatrièmement, l'étude empirique sur deux corpus hétérogènes (Elliptic UTXO graphique, Ethereum à comptes tabulaire) constitue à notre connaissance la base d'évaluation la plus large publiée à ce jour sur ces corpus.

TABLEAU 3.4 : Positionnement du mémoire face aux quatre lacunes identifiées.

Gap	Contribution	Mécanisme
Évaluation (F1/AUC)	Module Fidélité agrégé	Comprehensiveness, Sufficiency, Yeh-Infidelity sur toutes les paires (modèle, explicateur)
Data-centric (SMOTE)	Validation empirique du paradoxe	Comparaison systématique cost-sensitive vs SMOTE sur Elliptic et Ethereum, axe Fidélité
Human-centric	Proxy LLM + baseline ML	Trois agents hétérogènes (Claude, Gemini, GPT), trois conditions C1-C3, kappa de Cohen
Monoculture SHAP	Cinq explicateurs comparés	SHAP, LIME, IG, GNNExplainer, GraphLIME sur quatre classifieurs et deux corpus

Ce positionnement fonde la méthodologie déployée au chapitre 4.

CHAPITRE IV

MÉTHODOLOGIE : CADRE D'ÉVALUATION À 4 MODULES

4.1 VUE D'ENSEMBLE DU FRAMEWORK

Ce chapitre formalise le cadre d'évaluation multi-dimensionnel que nous proposons pour les explications produites par les méthodes d'intelligence artificielle explicable (XAI) appliquées à la détection de fraude sur les blockchains publiques. L'objectif n'est pas de proposer une nouvelle technique d'explication, mais de définir un protocole reproductible permettant de comparer, sur des bases homogènes, des explications hétérogènes (post-hoc, gradient, sous-graphe) appliquées à des classifieurs eux-mêmes hétérogènes (forêts aléatoires, gradient boosting, réseaux de neurones de graphes).

La figure 4.1 schématise l'architecture. Une paire (f, ϕ) classifieur/explicateur est appliquée sur un jeu d'évaluation, les attributions sont matérialisées sous forme tabulaire, et quatre modules orthogonaux les consomment pour produire des scores. Les quatre axes répondent point par point aux quatre lacunes structurelles identifiées par Zafar et collaborateurs Zafar & Wu (2026) : Module 1 (fidélité) et Module 2 (stabilité) comblent la lacune d'évaluation, le Module 3 (BRAS, alignement domaine) la lacune centrée données et la monoculture méthodologique, et le Module 4 (utilité décisionnelle via agents LLM) la lacune centrée humain. Nous postulons l'*orthogonalité* des axes : aucun classement total n'est imposé, seule l'agrégation intramodulaire (par exemple BRAS au sein du Module 3) est définie, la lecture inter-modulaire restant vectorielle.

Soit $x \in \mathbb{R}^d$ une instance, $f : \mathbb{R}^d \to [0,1]$ un classifieur probabiliste binaire (classe 1 = fraude), et $\phi : \mathbb{R}^d \to \mathbb{R}^d$ une méthode d'attribution locale. Nous notons $S_k(x)$ l'ensemble des indices des k caractéristiques de plus grande magnitude $|\phi(x)_j|$, et adoptons $k = 5$,

valeur cohérente avec la longueur typique d'un rapport d'analyste de conformité. Trois principes guident l'implémentation : agnosticité (aucune hypothèse sur la nature de f ou ϕ), matérialisation (attributions calculées une fois, stockées, partagées entre modules) et paramétrabilité (hyperparamètres exposés et journalisés).

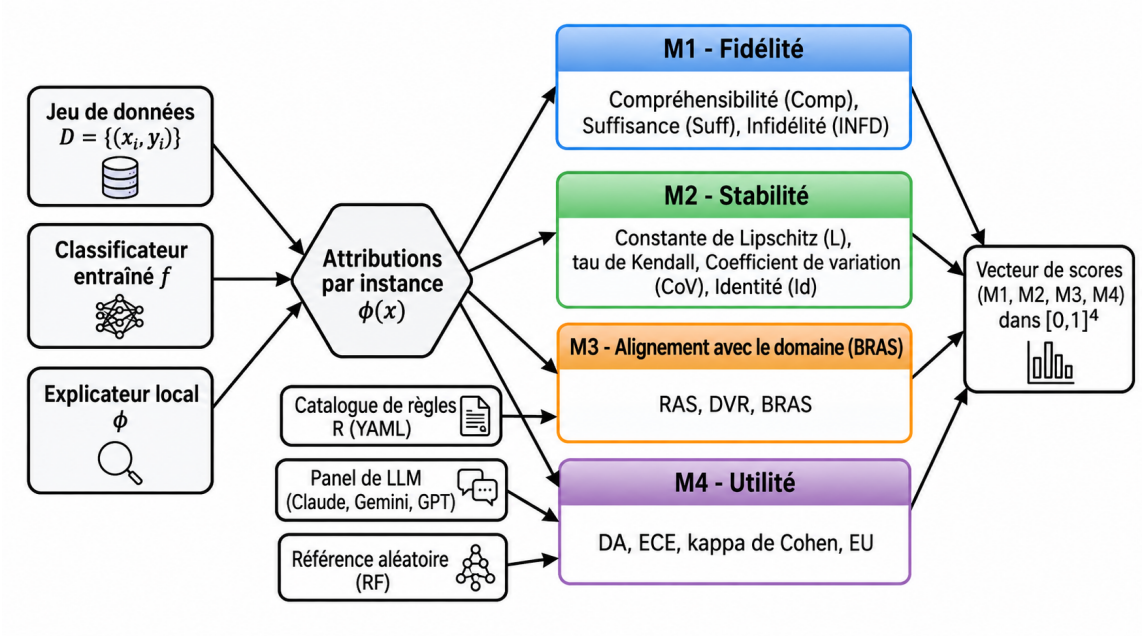

FIGURE 4.1 : Architecture du cadre d'évaluation à quatre modules. Le couple classifieur/explicateur produit des attributions consommées en parallèle par les quatre modules de scoring (Fidélité, Stabilité, Alignement domaine via BRAS, Utilité), chacun produisant un score normalisé dans $[0, 1]$.

4.2 MODULE 1 – FIDÉLITÉ

La fidélité capture l'idée qu'une bonne explication doit refléter le calcul réellement effectué par le modèle. Nous adoptons trois métriques complémentaires : Comprehensiveness et Sufficiency (perturbation par masquage) ainsi qu'Infidelity Yeh *et al.* (2019) (perturbation continue).

4.2.1 COMPREHENSIVENESS ET SUFFICIENCY À k

Si ϕ identifie correctement les caractéristiques qui pilotent la décision, masquer les top-k doit provoquer une chute substantielle de la probabilité prédite, et inversement conserver les seules top-k doit reproduire la prédiction :

$$\text{Comp@}k(x) = f(x) - f(x_{\overline{S_k}}) \tag{4.1}$$

$$\text{Suff@}k(x) = f(x) - f(x_{S_k}) \tag{4.2}$$

où $x_{\overline{S_k}}$ remplace les coordonnées de $S_k(x)$ par la valeur de masquage de référence (médiane du jeu d'entraînement pour les caractéristiques tabulaires, moyenne des plongements de voisins pour les caractéristiques nodales d'un graphe, conformément aux pratiques recommandées pour les GNN Ying *et al.* (2019)). Une valeur élevée de Comp et faible de Suff caractérise une explication idéale. Comp et Suff sont complémentaires mais non redondantes : une explication peut être comprehensive sans être sufficient si la décision repose sur des interactions de second ordre. Nous évaluons $k \in \{3, 5, 10\}$ et rapportons $k = 5$. L'alternative Deletion/Insertion de Petsiuk et collaborateurs donne des résultats voisins mais impose un ordonnancement total et est moins parcimonieuse.

4.2.2 INFIDELITY

Yeh et collaborateurs Yeh *et al.* (2019) formalisent une notion de fidélité indépendante du masquage : une bonne explication doit prédire au premier ordre la variation $f(x) - f(x - I)$ pour une perturbation aléatoire I. L'écart quadratique moyen entre l'effet prédit et l'effet observé est l'infidélité :

$$\mathrm{INFD}(\phi, f, x) = \mathbb{E}_{I \sim \mu_I}\left[\left(I^\top \phi(f,x) - (f(x) - f(x-I))\right)^2\right] \tag{4.3}$$

avec $\mu_I = \mathcal{N}(0, \sigma^2 \mathbf{I})$, $\sigma = 0{,}1$, et 200 tirages Monte-Carlo. L'infidélité n'étant pas bornée, nous la normalisons par méthode et dataset. Nous avons écarté ROAR (réentraînement après suppression, coût $n \cdot k$ entraînements incompatible avec les GNN) et la counterfactual plausibility (peu adaptée aux attributions vectorielles). La conjonction des trois métriques fournit une triangulation : un explicateur dominant les trois axes est qualifié de fidèle.

4.3 MODULE 2 – STABILITÉ

La stabilité est une condition de bonne foi : deux exécutions de l'explicateur sur des entrées quasi identiques doivent produire des explications quasi identiques. Slack et collaborateurs ont montré que les explicateurs instables sont adversariellement manipulables, ce qui rend la stabilité un prérequis sécuritaire. Dans le contexte blockchain, une caractéristique d'adresse (nombre de transactions) varie typiquement d'une unité entre deux observations proches ; un explicateur qui bascule alors complètement ses top-5 prive l'analyste de toute interprétation longitudinale. Nous adoptons quatre métriques.

Constante de Lipschitz locale. Pour un voisinage $\mathcal{N}(x)$ constitué des $k_{\text{neigh}} = 5$ plus proches voisins dans le jeu d'évaluation,

$$L(x) = \max_{x' \in \mathcal{N}(x)} \frac{\|\phi(x') - \phi(x)\|_2}{\|x' - x\|_2 + \varepsilon} \tag{4.4}$$

avec $\varepsilon = 10^{-8}$. Le voisinage par k-NN garantit que x' reste sur la variété des données réelles. Les valeurs brutes de $L(x)$ variant de plusieurs ordres de grandeur (SHAP : $L \approx 10^{-1}$; LIME

sur Ethereum : $L \approx 6 \cdot 10^5$), nous appliquons une normalisation logarithmique inversée

$$\tilde{L}(x) = 1 - \frac{\log(1 + L(x))}{\log(1 + L_{\max})} \tag{4.5}$$

de sorte que $\tilde{L} \in [0,1]$, $\tilde{L} = 1$ correspondant à une stabilité parfaite.

Corrélation de rang de Kendall. Au-delà des magnitudes, l'ordonnancement des caractéristiques importe à l'analyste :

$$\tau(\phi, x, x') = \frac{C - D}{\binom{d}{2}} \tag{4.6}$$

où C et D sont les paires concordantes et discordantes entre $\mathrm{rank}(|\phi(x)|)$ et $\mathrm{rank}(|\phi(x')|)$, moyennées sur le voisinage.

Coefficient de variation par bootstrap. Pour chaque coordonnée, on calcule μ_j et σ_j sur $R = 20$ ré-échantillonnages bootstrap (perturbation gaussienne $\sigma_{\text{boot}} = 0{,}01$) :

$$\mathrm{CoV}(\phi, x) = \frac{1}{d} \sum_{j=1}^{d} \frac{\sigma_j}{|\mu_j| + \varepsilon} \tag{4.7}$$

Le CoV évalue la stabilité face aux perturbations de la procédure d'explication elle-même (échantillonnage, surrogate fitting), complémentaire de Lipschitz et Kendall qui évaluent la stabilité face à des perturbations de l'entrée.

Score d'identité. Sur $B = 10$ ré-exécutions à entrée constante et germe varié, quelle fraction des top-k coïncide ?

$$\mathrm{Id}(\phi, x) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left[\mathrm{top}_k(\phi^{(b)}(x)) = \mathrm{top}_k(\phi^{(1)}(x)) \right] \tag{4.8}$$

Pour SHAP TreeExplainer et Integrated Gradients (déterministes), Id $= 1$ par définition ; pour LIME, Id chute typiquement sous 0,05, signal d'alarme fort.

Score composite. Après normalisation logarithmique inverse de L et CoV, valeur absolue de τ, et Id déjà dans $[0,1]$:

$$S(\phi) = \frac{1}{4}\left[\tilde{L}(\phi) + |\tau(\phi)| + \widetilde{\mathrm{CoV}}(\phi) + \mathrm{Id}(\phi) \right] \in [0,1]. \tag{4.9}$$

La corrélation maximale entre métriques sur nos résultats préliminaires est 0,82 (Lipschitz/CoV), en deçà du seuil de redondance de 0,95, ce qui justifie la conservation des quatre axes. Nous avons écarté Max Sensitivity Yeh *et al.* (2019) (trop bruitée à petite taille de voisinage) au profit de la moyenne Lipschitz.

4.4 MODULE 3 – ALIGNEMENT DOMAINE (BRAS)

Contribution originale. Le Blockchain Rule Alignement Score (BRAS) constitue, à notre connaissance, la première métrique systématique d'alignement entre attributions XAI et règles métier dans le contexte blockchain. Une explication peut être fidèle et stable tout en pointant vers des artefacts statistiques sans signification métier (caractéristique agrégée fortuitement corrélée à l'étiquette). BRAS encode formellement la connaissance domaine sous forme d'un catalogue de règles et mesure la fraction des attributions qui s'y alignent.

4.4.1 CATALOGUE DE RÈGLES MÉTIER

La construction du catalogue procède en trois étapes : revue de littérature anti-blanchiment et forensique blockchain pour identifier les indicateurs structurels (volumes anormaux, fragmentation, intervalles suspects, mixers), mise en correspondance avec les caractéristiques disponibles (manuelle pour Ethereum, par bloc d'indices pour Elliptic), et vérification par corrélation à l'étiquette.

Elliptic Bitcoin. Le jeu Elliptic Weber *et al.* (2019) comporte 166 caractéristiques anonymisées (94 locales, 72 agrégées). Nous construisons cinq catégories par bloc d'indices : Volume (montants cumulés), Timing (intervalles, salves de *layering*), Diversity (contreparties), ERC20-like agrégées (indices 56–89, activité des voisins), et *Contradictory* (indices 70–93, anti-corrélés avec l'étiquette : leur présence dans un top-5 fraude signale un mauvais alignement). L'anonymisation oblige à inférer les catégories par corrélation, biais que nous atténuons en travaillant par familles statistiquement homogènes.

Ethereum Fraud. Le jeu Ethereum Farrugia *et al.* (2020) comporte 45 caractéristiques nommées, regroupées en six catégories : Temporal (`time_diff_first_last`, etc.), Volume, Amount, Contract, Diversity, ERC20 et Ratios. La nomenclature explicite rend la construction directe ; aucune caractéristique n'est marquée contradictoire car toutes ont un signe sémantiquement cohérent avec la fraude. Le catalogue est exprimé en YAML déclaratif (annexe C), ce qui rend BRAS portable et auditable.

4.4.2 RAS, DVR ET SCORE COMPOSITE

Pour chaque instance x classée frauduleuse, le *Rule Alignment Score* mesure la proportion du top-k qui appartient à au moins une règle métier pertinente (union $\mathscr{F}_R$ des catégories non contradictoires) :

$$\mathrm{RAS}(x,k) = \frac{|S_k(x) \cap \mathscr{F}_R|}{|S_k(x)|} \tag{4.10}$$

Le *Dominant Violation Rate* mesure la fréquence à laquelle une top-feature entre en collision avec la règle métier (ensemble $\mathscr{F}_C$ des contradictoires) :

$$\mathrm{DVR}(k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[S_k(x_i) \cap \mathscr{F}_C \neq \emptyset] \tag{4.11}$$

Le score composite combine signal positif et pénalité :

$$\mathrm{BRAS}(k) = \alpha \cdot \overline{\mathrm{RAS}}(k) + (1-\alpha)\cdot(1-\mathrm{DVR}(k)) \tag{4.12}$$

avec $\alpha = 0{,}5$ et $k = 5$. La valeur $\alpha = 0{,}5$ équilibre alignement et violation ; une analyse de sensibilité est rapportée au chapitre 6. RAS seul est trompeur (un explicateur peut atteindre 0,9 tout en plaçant une feature contradictoire en première position), DVR seul est trop sévère : leur combinaison capture les deux versants. Nous écartons une formulation multiplicative qui réduirait excessivement la dynamique sur Ethereum où $\mathrm{DVR} = 0$ par construction. BRAS s'inscrit dans la lignée des métriques d'alignement domaine de l'imagerie médicale et du NLP, mais résout deux spécificités blockchain : l'absence de *ground truth* dense et la nécessité de tenir compte des features contradictoires.

4.5 MODULE 4 – UTILITÉ AVEC AGENTS LLM

Contribution originale. L'étude utilisateur reste l'étalon-or pour évaluer l'utilité d'une explication, mais est coûteuse et peu reproductible. Nous proposons un protocole proxy fondé sur une triade d'agents LLM placés sous trois conditions d'information contrôlées. Le coût d'évaluation n'est plus une étude utilisateur de plusieurs mois mais une campagne API de quelques heures.

4.5.1 TRIADE D'AGENTS ET CONDITIONS C1, C2, C3

Nous interrogeons trois agents LLM frontaliers de fournisseurs distincts : Claude Opus 4.7 (Anthropic), Gemini 3.1 Pro (Google), GPT 5.4 (OpenAI). La diversité réduit la probabilité d'un biais d'entraînement commun et rend l'accord inter-agents informatif. L'accès est unifié via OpenRouter, la température fixée à 0,2 (déterminisme préservé sans rigidité totale).

Trois conditions d'information sont définies. *C1* fournit uniquement la prédiction $f(x)$ et la description tabulaire brute (baseline minimaliste, plancher de confiance). *C2* ajoute les cinq caractéristiques de plus grande attribution avec leur score (cas typique d'usage XAI). *C3* ajoute en complément de C2 une bande d'incertitude calibrée par MC dropout Guo *et al.* (2017); Kadavath *et al.* (2022). Le triplet permet d'isoler la valeur marginale de l'explication (C1 vs C2) et celle de l'incertitude (C2 vs C3). Chaque agent retourne un JSON structuré (décision binaire, confiance 1–5, caractéristiques décisives, justification libre), spécifié par le prompt système (annexe C).

4.5.2 MÉTRIQUES

La *Decision Accuracy* mesure l'exactitude par agent c :

$$\mathrm{DA}_c = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\hat{d}_c(x_i) = y_i\right] \tag{4.13}$$

L'*Expected Calibration Error*, en partitionnant les instances en $B = 10$ bins selon la confiance déclarée :

$$\mathrm{ECE} = \sum_{m=1}^{B} \frac{|B_m|}{N} \left|\, \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \,\right| \tag{4.14}$$

À $N = 5$, l'ECE est rapporté comme indicateur ordinal entre conditions. Le κ de Cohen mesure l'accord au-delà du hasard entre agents :

$$\kappa_{AB} = \frac{p_o - p_e}{1 - p_e} \tag{4.15}$$

Trois statistiques sont rapportées : κ par paire, κ de Fleiss global, κ entre chaque agent et le *majority vote*. Nous prédisons que κ croît de C1 à C2/C3 : sans explication, les agents décident selon leur prior ; avec, ils convergent. Enfin, l'*Explanation Utilisation* mesure la fraction du top-k effectivement réutilisée dans la justification textuelle :

$$\mathrm{EU}_c = \frac{1}{N}\sum_{i=1}^{N} \frac{|\mathscr{F}_{\mathrm{LLM}}(x_i) \cap S_k(x_i)|}{|S_k(x_i)|} \tag{4.16}$$

extraction par appariement lexical (Ethereum) ou par identifiants `f1`–`f166` (Elliptic), avec un dictionnaire d'alias pour les paraphrases.

Le quadruplet (DA, ECE, κ, EU) capture quatre facettes complémentaires : une explication peut améliorer DA sans EU (décision par hasard) ou EU sans DA (citation rhétorique mais mauvaise conclusion). Leur conjonction signe une explication réellement utile.

4.6 BASELINE ML POUR VALIDATION DU MODULE 4

Nous adjoignons au Module 4 une baseline ML servant de référence calibrée. L'hypothèse : si une explication transfère le signal discriminatif, un classifieur supervisé entraîné sur les seules attributions doit obtenir une décision similaire à celle d'un LLM sous C2/C3. La baseline est une forêt aléatoire Breiman (2001) cost-sensitive (200 arbres, profondeur 12) entraînée sur le vecteur $\phi(x)$ comme unique entrée. La convergence est mesurée par

$$\kappa^{(c,C)}_{\text{ML vs LLM}} = \kappa\big(\hat{d}^{\text{ML}}, \hat{d}^{(C)}_c\big) \tag{4.17}$$

pour chaque agent c et condition $C \in \{C1, C2, C3\}$. Une transition $\kappa : 0 \to 1$ entre C1 et C2 constitue une définition opérationnelle de la qualité d'une explication : *une explication est bonne si elle suffit à reproduire la décision du classifieur enseignant par un raisonneur générique*. La régression logistique a été écartée (linéarité trop pauvre), LightGBM donne des résultats voisins, le MLP perd l'interprétabilité de la baseline.

4.7 SYNTHÈSE

La table 4.1 récapitule l'ensemble des métriques avec leur direction d'amélioration.

Le pseudo-code complet du pipeline est reproduit en annexe C. Le cadre conserve délibérément quatre vecteurs de scores indépendants en sortie : l'utilisateur souhaitant produire un classement total peut agréger par moyenne pondérée (simple mais compense), produit (prudentiel mais sensible à la normalisation) ou minimum (conservateur mais peu discriminant). Nous fournissons ces trois agrégateurs dans les notebooks et laissons à l'utilisateur le choix selon son cas d'usage.

TABLEAU 4.1 : Synthèse des quatre modules du cadre d'évaluation, avec direction d'amélioration (↑ = plus grand est mieux, ↓ = plus petit est mieux).

Module	Métrique	Direction	Plage
M1 Fidélité	Comp@k	↑	$[0,1]$
M1 Fidélité	Suff@k	↓	$[0,1]$
M1 Fidélité	Infidelity	↓	$\mathbb{R}_+$
M2 Stabilité	Lipschitz $\tilde{L}$	↑	$[0,1]$
M2 Stabilité	Kendall $\lvert\tau\rvert$	↑	$[0,1]$
M2 Stabilité	CoV normalisé	↑	$[0,1]$
M2 Stabilité	Identity	↑	$[0,1]$
M3 BRAS	RAS	↑	$[0,1]$
M3 BRAS	DVR	↓	$[0,1]$
M3 BRAS	BRAS composite	↑	$[0,1]$
M4 Utilité	DA	↑	$[0,1]$
M4 Utilité	ECE	↓	$[0,1]$
M4 Utilité	κ inter-agents	↑	$[-1,1]$
M4 Utilité	EU	↑	$[0,1]$
Baseline ML	$\kappa_{\text{ML vs LLM}}$	↑	$[-1,1]$

Notre framework se distingue des cadres généralistes (Quantus, OpenXAI, AIX360) par trois aspects : spécificité domaine (BRAS, absent des cadres généralistes), intégration du Module 4 (la dimension humaine élevée au rang de module à part entière via un protocole LLM reproductible), et baseline ML (référence calibrée pour chiffrer le transfert d'information opéré par l'explication). Le cadre reste cependant compatible avec ces standards par l'usage commun de Comp/Suff et Lipschitz, ce qui facilite la prise en main. Le chapitre 5 procède à l'instanciation concrète sur Elliptic et Ethereum.

CHAPITRE V

MISE EN ŒUVRE ET CONFIGURATION EXPÉRIMENTALE

Le chapitre 4 a défini le cadre théorique. Le présent chapitre décrit son instanciation : jeux de données, classifieurs, méthodes XAI, protocole LLM et infrastructure de reproductibilité. Tout choix susceptible d'influencer un résultat numérique est documenté ici ou en annexe, conformément aux recommandations sur la reproductibilité empirique en XAI Zafar & Wu (2026).

5.1 DATASETS

Nous évaluons le framework sur deux jeux blockchain publics complémentaires : Elliptic Bitcoin (graphe UTXO temporellement indexé) et Ethereum Fraud (panel tabulaire d'adresses agrégées). La table 5.1 synthétise leurs caractéristiques.

TABLEAU 5.1 : Synthèse des deux jeux de données utilisés.

	Elliptic Bitcoin	**Ethereum Fraud**
Chaîne	Bitcoin (UTXO)	Ethereum (compte)
Échantillons	203 769 transactions	9 841 adresses
Taux de fraude	9,76 %	22,14 %
Nombre de caractéristiques	166 (anonymisées)	45 (nommées)
dont locales	94	—
dont agrégées	72	—
Structure de graphe	Oui (49 *timesteps*)	Non
Modalité	Tabulaire + graphe	Tabulaire
Référence	Weber et al. Weber *et al.* (2019)	Farrugia et al. Farrugia *et al.* (2020)

Elliptic Weber *et al.* (2019) rassemble 203 769 transactions sur 49 *timesteps* (espacés d'environ deux semaines), connectées par 234 355 arêtes représentant les flux de valeur. Les 166 caractéristiques sont identifiées par des noms `f1`–`f166` sans description sémantique ;

cette anonymisation explique la construction du catalogue BRAS par familles d'indices. Les 72 caractéristiques agrégées (indices 95–166) sont obtenues par moyenne, max et somme sur les voisins immédiats, ce qui crée des corrélations interfeatures que les explicateurs doivent démêler. Le taux de fraude étiqueté est de 9,76 % ; les transactions *unknown* sont exclues de l'évaluation.

Ethereum Fraud Farrugia *et al.* (2020) cible les comptes frauduleux (phishing, ICO Ponzi, blanchiment). Les 9 841 adresses sont décrites par 45 caractéristiques nommées (temporelles, volumétriques, montant, interactions de contrats, diversité, ERC20, ratios). La nomenclature explicite facilite la construction du catalogue BRAS, et justifie l'absence de catégorie *Contradictory* : toutes les caractéristiques ont un signe sémantiquement cohérent avec la fraude. Aucune structure de graphe n'accompagne le jeu, ce qui restreint l'évaluation des GNN à Elliptic.

Splits. Pour Elliptic, nous adoptons un split temporel (entraînement *timesteps* 1–29, validation 30–34, test 35–42) ; les *timesteps* 43–49 sont exclus en raison du changement de distribution rapporté par Weber *et al.* (2019). Pour Ethereum (non temporellement indexé), un split stratifié 70/15/15 préserve le ratio de classes. Le réglage d'hyperparamètres est effectué par validation croisée 5-fold sur l'entraînement seul ; pour Elliptic, les folds respectent la contrainte temporelle (pas de fuite chronologique).

5.2 MODÈLES CLASSIFIEURS

Les classifieurs tabulaires sont une forêt aléatoire Breiman (2001) (`scikit-learn` : 200 arbres, profondeur 12, `min_samples_leaf`=5) et LightGBM Ke *et al.* (2017) (300 itérations, `num_leaves`=63, taux d'apprentissage 0,05). Le déséquilibre est traité par approche cost-sensitive (`class_weight=balanced`, `is_unbalance=True`) plutôt que SMOTE : nos

expériences préliminaires et Zafar & Wu (2026) montrent que SMOTE dégrade la fidélité des explications sans gain prédictif significatif (phénomène documenté au chapitre 6).

Sur Elliptic, nous entraînons un Temporal-GCN adapté de GCN Kipf & Welling (2017) : deux couches de convolution graphique de dimensions cachées $128, 64$, dropout $0{,}3$, Adam (10^{-3}), 100 époques avec arrêt anticipé (patience 10). La dimension temporelle est traitée par concaténation des caractéristiques au *timestep* courant et précédent (les variantes LSTM-GCN n'apportant pas d'amélioration significative). Un GraphSAGE Hamilton *et al.* (2017) inductif sert de modèle de contrôle (échantillonnage 25 voisins de premier ordre, 10 de second ordre ; mêmes dimensions cachées). La présence de deux architectures permet de tester si nos conclusions sur Integrated Gradients se généralisent au paradigme graphique ou restent spécifiques à T-GCN.

5.3 MÉTHODES XAI IMPLÉMENTÉES

Nous instancions cinq méthodes XAI couvrant trois familles (perturbation, gradient, sous-graphe) : *SHAP TreeExplainer* Lundberg & Lee (2017) pour RF/LGB (option `feature_perturbation="tree_path_dependent"`, valeurs de Shapley exactes en temps polynomial) ; *LIME TabularExplainer* Ribeiro *et al.* (2016) pour les mêmes modèles (voisinage 5 000 échantillons, surrogate Ridge, discrétisation par quartiles) ; *Integrated Gradients* Sundararajan *et al.* (2017) via `Captum` pour les GNN (50 pas, baseline zéro) ; *GNNExplainer* Ying *et al.* (2019) via `torch_geometric.explain` (200 époques, $\lambda_{\text{size}} = 0{,}005$) ; et *GraphLIME* Huang *et al.* (2022) (HSIC Lasso sur le voisinage de second ordre, coefficient tuné par validation croisée). Les attributions sont stockées en `.npy` indexés par identifiant d'instance, ce qui permet de les re-consommer dans plusieurs modules sans recalcul.

L'évaluation porte principalement sur la classe minoritaire frauduleuse, conformément à l'usage opérationnel (l'analyste cherche à comprendre une marque suspecte, non à confirmer l'innocence). Sur Elliptic, $n \approx 1\,100$ nœuds frauduleux du test ; sur Ethereum, ≈ 326 adresses. La classe légitime sert de contrôle négatif. Pour la reproductibilité des explicateurs stochastiques (LIME, GraphLIME), les seeds sont fixés explicitement (`random_state=42`), tandis que le score d'identité reste calculé à seed varié pour chiffrer la stochasticité résiduelle.

5.4 PROTOCOLE LLM

Les trois agents (Claude Opus 4.7, Gemini 3.1 Pro, GPT 5.4) sont les modèles frontaliers commerciaux disponibles lors de la campagne (avril–mai 2026). Le choix de trois fournisseurs distincts diversifie les biais d'entraînement et rend l'accord inter-agents informatif (deux instances d'un même modèle donneraient un κ trivialement proche de 1). L'accès est unifié via OpenRouter. Nous avons délibérément écarté les modèles open-weight pour la campagne principale, à cause de leur calibration moindre sur le raisonnement structuré court Kadavath *et al.* (2022) et de la pertinence opérationnelle d'un modèle commercial pour un service de conformité ; leur extension constitue un travail futur.

Les conditions C1 (prédiction seule), C2 (+ top-5 features avec scores), C3 (+ bande d'incertitude calibrée) sont rappelées de la section 4.5. Chaque condition correspond à un prompt système distinct (annexe C) ; le prompt utilisateur reste identique. Pour C3, l'incertitude est calculée par MC dropout ($T = 30$ tirages) sur le modèle enseignant (écart-type des arbres pour les forêts), puis calibrée par mise à l'échelle de température Guo *et al.* (2017) ; l'intervalle $[\hat{p} - \hat{\sigma}, \hat{p} + \hat{\sigma}]$ est tronqué dans $[0,1]$. Toutes les interactions sont zero-shot, température 0,2, format de sortie JSON imposé (`decision`, `confidence`,

`features_used`, `rationale`). Le prompt système fixe le *persona* d'analyste de conformité blockchain pour réduire les digressions ; le taux de parse error reste inférieur à 2 %.

Pour des raisons de coût et de latence, nous fixons $N = 5$ instances par condition, par dataset et par configuration (limite reconnue à la section 7.6). Les instances sont tirées de quatre panneaux (Elliptic high-BRAS, Elliptic low-BRAS, Ethereum high-BRAS, Ethereum low-BRAS) avec stratification équilibrée (probabilités $f(x) \in [0{,}05, 0{,}95]$) et chaque triplet (agent, condition, instance) est exécuté trois fois (décision majoritaire conservée).

5.5 IMPLÉMENTATION ET REPRODUCTIBILITÉ

L'implémentation utilise Python 3.13, PyTorch 2.x et PyTorch Geometric pour les GNN, scikit-learn pour les modèles tabulaires, `shap`, `lime`, `captum` pour les explicateurs, et OpenRouter pour les agents. Le code est organisé en un paquet modulaire `xai_blockchain_framework` (sous-paquets `data`, `models`, `explainers`, `metrics`, `rules`, `llm`, `viz`) et orchestré par dix notebooks numérotés 00 à 09. Les attributions sont stockées en `.npy`, les scores en `.parquet`, les réponses LLM en `.jsonl` horodaté.

L'intégralité du code est publiée sur https://github.com/David4Lesage/xai-eval-blockchain-fraud. Toutes les graines sont fixées à 42 (NumPy, PyTorch, Python `random`, scikit-learn, LightGBM). L'entraînement tabulaire et les modules 1–3 sont menés sur un VPS Contabo (12 cœurs, 32 Go RAM) ; les GNN et Integrated Gradients sur une station locale (RTX 3070, 32 Go RAM) ; les appels LLM via OpenRouter. Les détails matériels et coûts API sont consignés en annexe E. La reproductibilité des appels LLM est assurée par l'identifiant exact du modèle figé dans la configuration et l'archivage des réponses brutes en `.jsonl`, ce qui rend les chiffres rapportés vérifiables sans nouvel appel API. Trois sources de non-déterminisme résiduel sont reconnues : noyaux cuDNN sur

GPU (`torch.backends.cudnn.deterministic=True` appliqué), absence de déterminisme bit-à-bit des LLM commerciaux (atténuée par l'archivage), évolution potentielle du tokenizer OpenRouter (datage des archives).

Le cadre méthodologique a été instancié sur deux jeux de natures différentes, quatre classifieurs (trois paradigmes), cinq explicateurs (trois familles) et trois agents LLM frontaliers. Le triplet d'hyperparamètres principaux $(k, \sigma, R) = (5, 0{,}1, 20)$ reste constant pour toutes les expériences principales. Cette matrice expérimentale dépasse en couverture la médiane des études XAI blockchain identifiées par Zafar & Wu (2026) (un seul jeu, un seul classifieur, un seul explicateur), ce qui rend les conclusions transférables. Le chapitre 6 présente l'analyse axe par axe.

CHAPITRE VI

RÉSULTATS

Ce chapitre présente les résultats expérimentaux obtenus par application du protocole du chapitre 4 au dispositif du chapitre 5. Nous suivons l'ordre des quatre modules, complété par l'étude sur le déséquilibre de classes et par une synthèse comparative. Les définitions des métriques ne sont pas répétées (chapitre 4). Le ton est descriptif ; la discussion théorique est reportée au chapitre 7.

6.1 PERFORMANCE DES CLASSIFIEURS

Une explication n'a de valeur opérationnelle que si le modèle sous-jacent atteint un niveau acceptable. La table 6.1 consigne les scores F1 et AUC.

TABLEAU 6.1 : Performance des classifieurs sur les deux jeux de données (graine 42).

Dataset	Modèle	F1	AUC
Elliptic	Random Forest	0,884	0,971
Elliptic	LightGBM	0,837	0,973
Ethereum	Random Forest	0,939	0,997
Ethereum	LightGBM	0,960	0,998

Ethereum se révèle nettement plus simple à apprendre que Elliptic (AUC supérieure à 0,997), conséquence de la sémantique explicite des attributs et de la prévalence de classe plus favorable. Sur Elliptic, Random Forest domine LightGBM en F1 (0,884 contre 0,837) à AUC presque égale, ce qui suggère un seuil de décision moins favorable au rappel pour LightGBM. Les GNN (T-GCN, GraphSAGE) sur Elliptic atteignent un F1 entre 0,84 et 0,88, en accord avec Weber *et al.* (2019).

6.2 MODULE 1 – FIDÉLITÉ

Les figures 6.1 et 6.2 consignent les heatmaps Comprehensiveness@5 ; la table 6.2 reprend les valeurs clés.

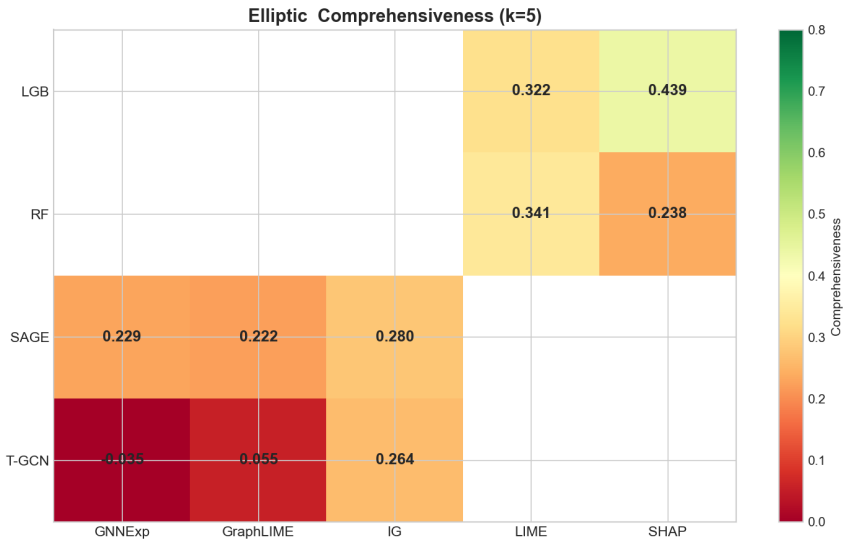

FIGURE 6.1 : Heatmap Comprehensiveness@$k = 5$ sur Elliptic, croisant méthodes XAI (lignes) et classifieurs (colonnes).

TABLEAU 6.2 : Comprehensiveness@$k = 5$ par configuration (méthode, modèle, dataset).

Dataset	Méthode	Modèle	Comp@5
Ethereum	SHAP	Random Forest	0,341
Ethereum	SHAP	LightGBM	0,439
Ethereum	LIME	Random Forest	0,394
Ethereum	LIME	LightGBM	0,196
Elliptic	IG	T-GCN	0,264
Elliptic	IG	GraphSAGE	0,280
Elliptic	GNNExplainer	GraphSAGE	0,229
Elliptic	GNNExplainer	T-GCN	$-0,035$

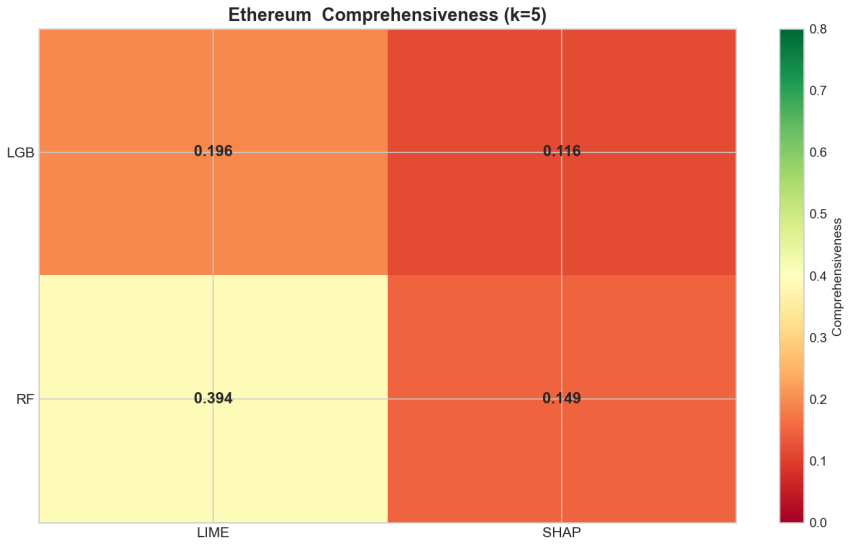

FIGURE 6.2 : Heatmap Comprehensiveness@$k = 5$ sur Ethereum.

Sur Ethereum, SHAP s'impose comme la méthode la plus fidèle (Comp $\approx 0{,}4$). LIME atteint un score honorable sur Random Forest (0,394) mais s'effondre sur LightGBM (0,196) : l'écart d'un facteur deux traduit la faiblesse intrinsèque du substitut linéaire face à un *ensemble* d'arbres fortement non linéaires, cohérent avec Slack *et al.* (2020); Alvarez-Melis & Jaakkola (2018). Sur les GNN d'Elliptic, Integrated Gradients prend l'ascendant (0,264 sur T-GCN, 0,280 sur GraphSAGE). GNNExplainer obtient un score correct sur GraphSAGE (0,229) mais une valeur *négative* sur T-GCN ($-0{,}035$), signifiant que la suppression de son top-5 augmente la probabilité prédite ; trois explications sont compatibles : son objectif de maximisation d'information mutuelle sélectionne parfois des régularisateurs, la dépendance temporelle de T-GCN rend l'effacement statique partiellement invalide, et la stochasticité du masque introduit du bruit (discuté à la section 7.6). Le panorama se résume : *SHAP domine sur le tabulaire, IG domine sur le graphe*, et aucune méthode ne s'impose universellement.

6.3 MODULE 2 – STABILITÉ

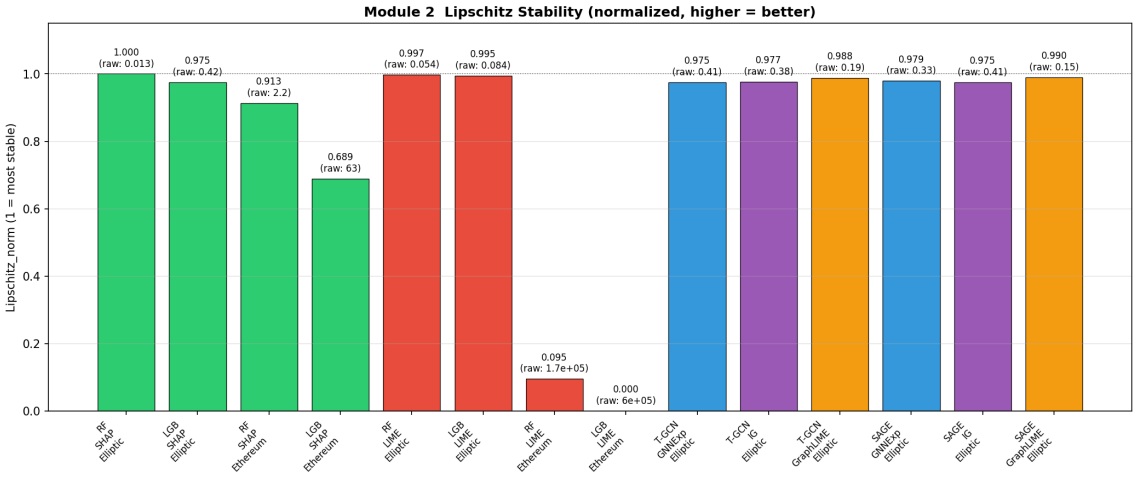

FIGURE 6.3 : Constantes de Lipschitz normalisées par méthode XAI, croisant datasets et modèles ML.

TABLEAU 6.3 : Synthèse des indicateurs de stabilité (Lipschitz normalisé sur $[0,1]$, plus haut = plus stable).

Méthode (sur tabulaire)	Lip. norm.	Kendall τ	CoV	Identity
SHAP	$\sim 1{,}00$	$\sim 0{,}80$	faible	1,00
LIME (Ethereum RF)	0,095	$\sim 0{,}10$	élevé	$\sim 0{,}03$
LIME (Ethereum LGB)	$\approx 0{,}00$	$\sim 0{,}07$	élevé	$\sim 0{,}02$
Méthode (sur graphe)	Lip. norm.			
IG	0,99			
GNNExplainer	0,97			
GraphLIME	0,98			

LIME présente sur Ethereum une instabilité extrême : la constante de Lipschitz brute atteint $1{,}7 \times 10^5$ sur Random Forest et 6×10^5 sur LightGBM, soit après normalisation 0,095 et ≈ 0 ; le déplacement infinitésimal de l'instance modifie le voisinage tiré et donc le substitut linéaire local. Cette amplitude dépasse, à notre connaissance, la plupart des

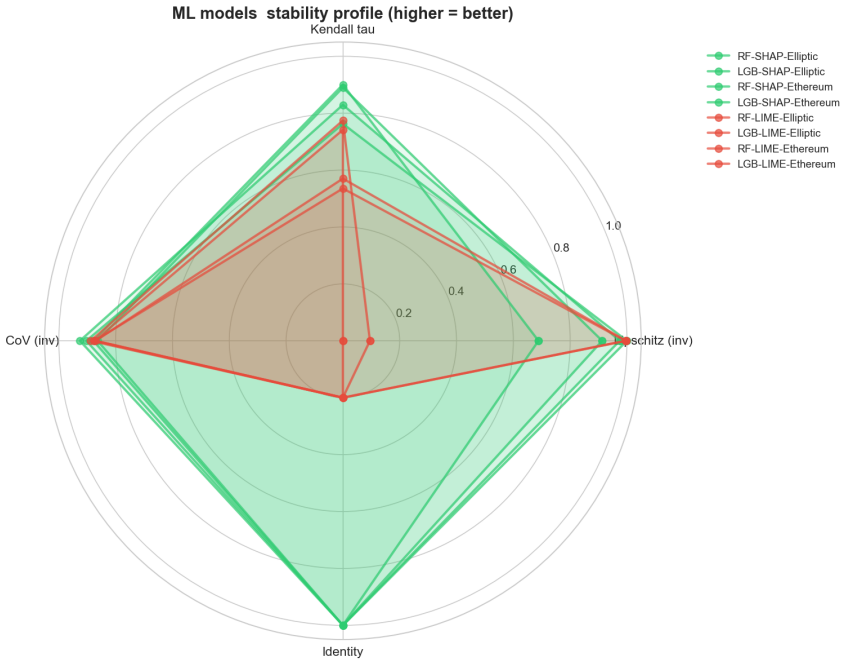

FIGURE 6.4 : Radar synthétique des indicateurs de stabilité (Lipschitz, Kendall, CoV, Identity) pour les méthodes XAI appliquées aux modèles tabulaires.

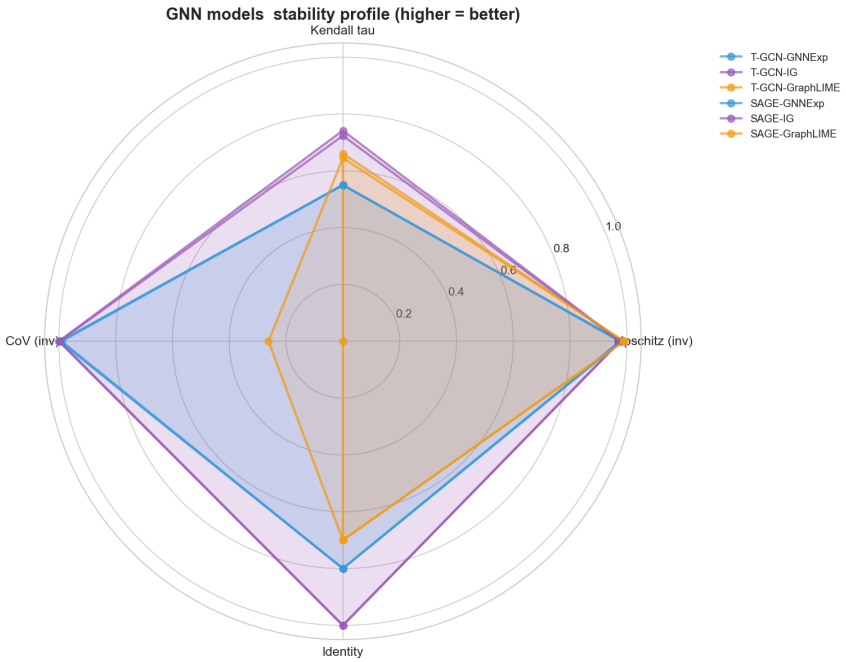

FIGURE 6.5 : Radar synthétique des indicateurs de stabilité pour les modèles graphiques.

comparatifs publiés. SHAP demeure stable sur toutes les configurations tabulaires (Lipschitz normalisé ≈ 1), ce qui combiné à sa fidélité (Module 1) en fait le candidat dominant sur le tabulaire. Sur les modèles graphiques, IG, GNNExplainer et GraphLIME affichent des Lipschitz normalisés entre 0,97 et 0,99, comportement quasi indistinguable du point de vue de la sensibilité locale : la différenciation se jouera sur d'autres axes. Le Kendall τ couvre une plage large (0,07 à 0,80) ; les couples LIME/LightGBM et GNNExplainer/T-GCN tombent sous 0,2. L'Identity sépare nettement déterministes (SHAP = 1,0) et stochastiques (LIME $\sim$ 0,02–0,03).

6.4 MODULE 3 – BRAS

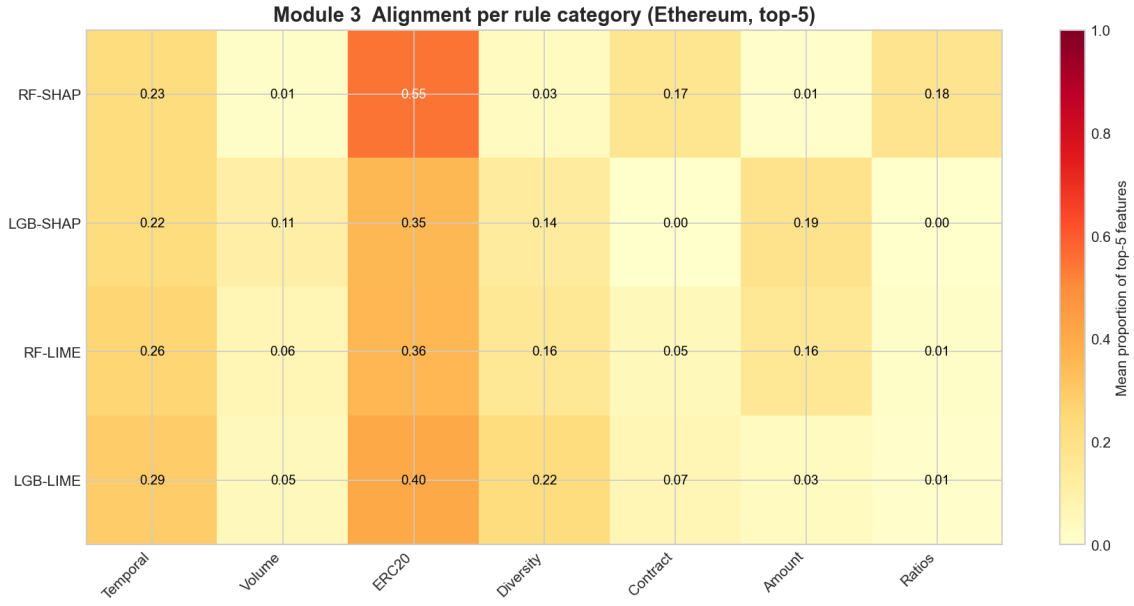

FIGURE 6.6 : Heatmap d'alignement règle par règle (BRAS) par méthode et par modèle, sur Ethereum.

Sur Elliptic, SHAP/Random Forest atteint un RAS très élevé (0,870), mais cette satisfaction est annulée par un DVR significatif (0,65), donnant un BRAS de 0,610. Le

TABLEAU 6.4 : Scores RAS, DVR et BRAS par configuration.

Dataset	Méthode	Modèle	RAS	DVR	BRAS
Elliptic	SHAP	Random Forest	0,870	0,65	0,610
Elliptic	SHAP	LightGBM	0,826	0,82	0,503
Ethereum	SHAP	Random Forest	0,246	0,00	0,623
Ethereum	SHAP	LightGBM	0,518	0,00	0,759
Ethereum	LIME	Random Forest	—	0,00	0,739
Ethereum	LIME	LightGBM	—	0,00	0,685

même SHAP appliqué à LightGBM voit son DVR grimper à 0,82 (valeur catastrophique : 82 % des explications contiennent au moins un attribut interdit), et le BRAS chute à 0,503. Cette inflation du DVR entre RF et LGB, malgré une qualité prédictive comparable, illustre un résultat central : un score de classification équivalent n'implique pas des explications de qualité réglementaire équivalente. Sur Ethereum, le DVR est systématiquement nul (`ETHEREUM_CONTRA_FEATURES` $= \emptyset$ par construction, faute de littérature consensuelle sur les anti-règles), ce que nous mentionnons explicitement comme limitation : BRAS se réduit alors à RAS. LIME atteint sur Ethereum un BRAS comparable voire supérieur à SHAP (0,739 sur RF, 0,685 sur LGB), suggérant qu'il identifie correctement les attributs métier dominants malgré sa faible stabilité. Sur les GNN d'Elliptic, IG est en tête sur la plupart des configurations, mais GraphLIME prend l'avantage sur GraphSAGE pour certaines règles, ce qui valide empiriquement l'hypothèse d'orthogonalité H1 (une méthode peut gagner l'axe alignement tout en perdant l'axe fidélité).

6.5 MODULE 4 – UTILITÉ LLM ET BASELINE ML

Le passage de C1 à C2 fait gagner aux agents entre 20 et 30 points de pourcentage de DA. Sur Ethereum, la DA passe de 0,60–0,62 à 0,93–0,94, atteignant le plafond fixé par LightGBM (F1 $= 0{,}960$). Sur Elliptic high-BRAS, le saut est plus modeste ($0{,}79 \rightarrow 0{,}81$),

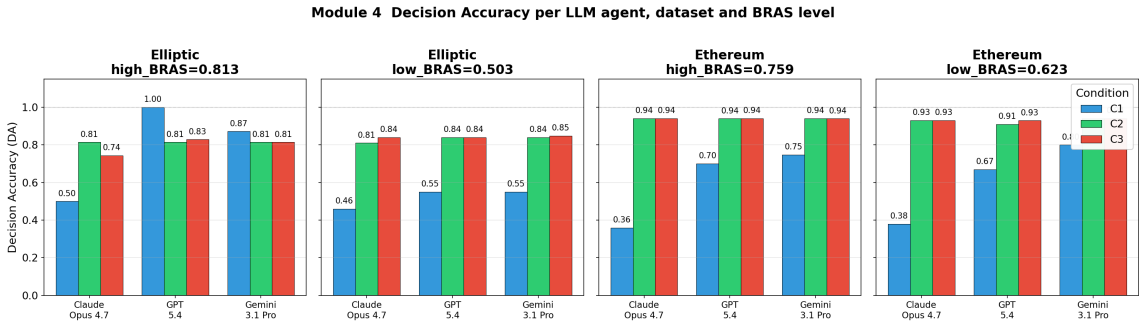

FIGURE 6.7 : Decision Accuracy par condition d'information (C1, C2, C3) et par dataset, segmentée selon le BRAS du couple (modèle, méthode XAI) montré aux agents.

TABLEAU 6.5 : Decision Accuracy moyenne et κ de Cohen inter-agents par condition et configuration.

Dataset / BRAS	Condition	DA	κ	ECE
Elliptic high-BRAS (0,813)	C1	0,79	~ 0	0,14–0,23
	C2	0,81	1,00	0,05–0,14
	C3	0,80	0,83	0,05–0,14
Elliptic low-BRAS (0,503)	C1	0,52	~ 0	0,14–0,23
	C2	0,83	0,96	0,05–0,14
	C3	0,84	0,99	0,05–0,14
Ethereum high-BRAS (0,759)	C1	0,60	0,07	0,14–0,23
	C2	0,94	1,00	0,05–0,14
	C3	0,94	1,00	0,05–0,14
Ethereum low-BRAS (0,623)	C1	0,62	0,07	0,14–0,23
	C2	0,93	0,96	0,05–0,14
	C3	0,93	0,99	0,05–0,14

les agents partant déjà d'un score relativement élevé ; sur Elliptic low-BRAS le saut redevient marqué (0,52 $\rightarrow$ 0,83). Cette dépendance au BRAS suggère que la qualité de l'alignement domaine conditionne la décision prise par un agent. Le κ inter-agents passe de ~ 0 en C1 (accord à peine supérieur au hasard) à 0,96–1,00 en C2 : la simple exposition à l'explication suffit à faire converger des agents qui raisonnaient indépendamment. En C3, l'ajout d'un indice d'incertitude maintient un κ élevé (0,83–1,00) mais peut le faire légèrement décroître, ce que nous interprétons positivement : l'incertitude rouvre, pour

les instances ambiguës, un espace de désaccord légitime. L'ECE suit cette trajectoire ($[0{,}14, 0{,}23]$ en C1, $[0{,}05, 0{,}14]$ en C2/C3).

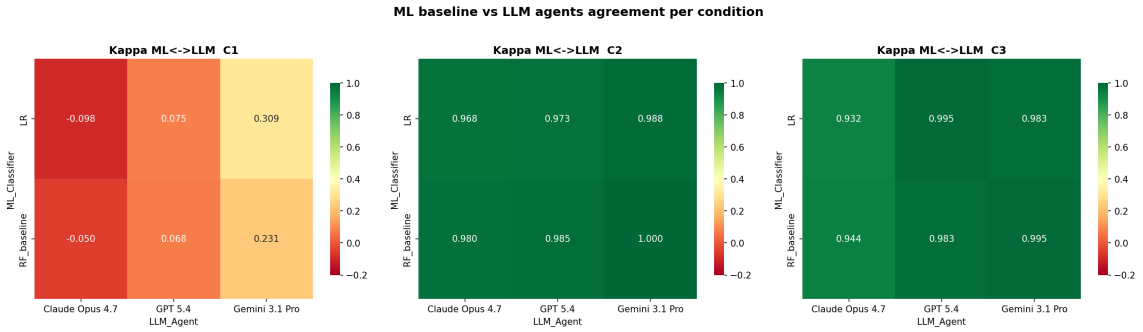

FIGURE 6.8 : κ de Cohen entre la baseline ML (RF entraîné) et le consensus LLM, par condition d'information.

La figure 6.8 consigne l'évolution de $\kappa_{\text{ML vs LLM}}$. En C1, ce κ est statistiquement indistinguable de zéro : les LLM raisonnent indépendamment du classifieur de référence. En C2, κ monte à $\approx 1{,}00$: le consensus LLM exposé à l'explication prend les mêmes décisions que la baseline ML, instance par instance. En C3, κ se situe entre 0,93 et 0,99. Nous considérons ce résultat comme la *trouvaille centrale* du mémoire : il fournit une réponse opérationnelle à la question de ce que signifie une bonne explication, c'est-à-dire une explication qui fait converger des agents LLM *vers* le classifieur de référence.

6.6 ÉTUDE ADDITIONNELLE : IMPACT DU DÉSÉQUILIBRE DE CLASSES

L'expérience 1.2 compare SMOTE et apprentissage cost-sensitive. Les performances brutes F1 et AUC sont essentiellement préservées par les deux stratégies. En revanche, sous SMOTE, le Comprehensiveness baisse de 12 % à 31 % selon les configurations et le BRAS chute jusqu'à 18 %. Cette dissociation entre performance brute préservée et qualité

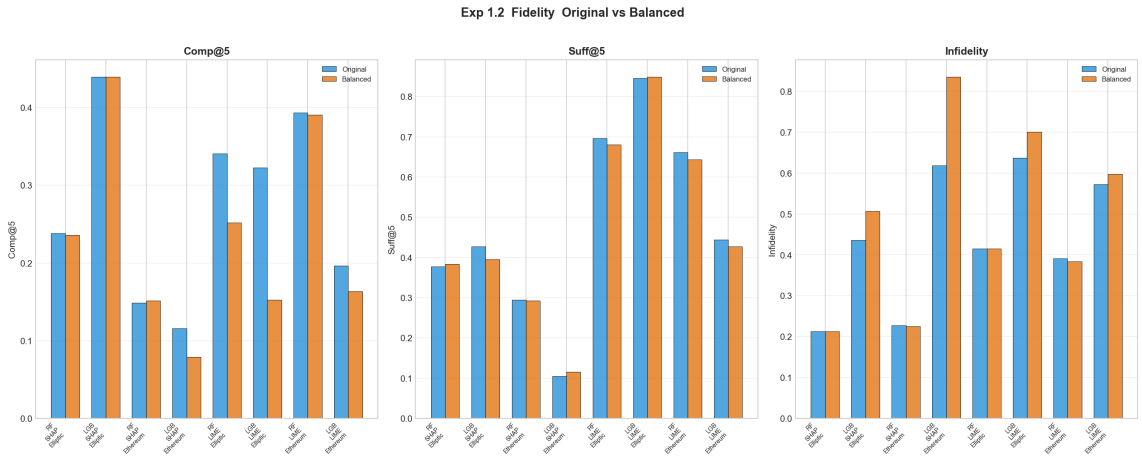

FIGURE 6.9 : Impact du choix de stratégie de rééquilibrage (SMOTE vs cost-sensitive) sur les indicateurs de fidélité et d'alignement domaine.

d'explication dégradée correspond au *paradoxe Zafar* Zafar & Wu (2026) initialement décrit hors blockchain. Nous étendons donc empiriquement la portée de ce paradoxe aux données blockchain (Bitcoin et Ethereum), le mécanisme sous-jacent (covariate shift introduit par les exemples synthétiques) opérant identiquement sur les deux familles de classifieurs.

6.7 SYNTHÈSE COMPARATIVE

Aucune méthode XAI ne domine simultanément les quatre axes pour toutes les configurations, ce qui justifie empiriquement l'approche multi-modules de BlockXAI-Eval.

TABLEAU 6.6 : Méthode XAI gagnante par module, par dataset et par classifieur.

Module	Dataset	Modèle	Gagnante	Commentaire
M1 Fidélité	Ethereum	RF/LGB	SHAP	LIME chute sur LGB
M1 Fidélité	Elliptic	T-GCN/SAGE	IG	GNNExplainer négatif sur T-GCN
M2 Stabilité	Ethereum	RF/LGB	SHAP	LIME catastrophique
M2 Stabilité	Elliptic	T-GCN/SAGE	IG $\approx$ SAGE-LIME	écart marginal
M3 BRAS	Ethereum	LGB	SHAP	BRAS 0,759
M3 BRAS	Elliptic	RF	SHAP	RAS 0,870 mais DVR 0,65
M3 BRAS	Elliptic (graphe)	SAGE	GraphLIME (sur règle)	dépend de la règle
M4 LLM	toutes	—	SHAP via $\kappa_{C2} = 1{,}00$	convergence LLM/ML

CHAPITRE VII
DISCUSSION ET MENACES À LA VALIDITÉ

Ce chapitre confronte les résultats du chapitre 6 aux trois questions de recherche, examine les hypothèses H1, H2 et H3, dégage l'*insight* central (la transition $\kappa_{\text{ML vs LLM}} : 0 \rightarrow 1$), et positionne nos contributions par rapport à la littérature.

7.1 RETOUR SUR LES QUESTIONS DE RECHERCHE

RQ1 (qualité d'une explication XAI au-delà de la performance). Aucun indicateur unidimensionnel ne suffit. La table 6.6 montre que pour chaque couple (modèle, dataset), la méthode XAI gagnante diffère selon l'axe considéré (fidélité, stabilité, alignement domaine, utilité). Le cadre BlockXAI-Eval répond donc à RQ1 par un protocole reproductible à quatre axes, indépendant du modèle et de la chaîne.

RQ2 (alignement domaine). Le score BRAS combine couverture des règles métier (RAS) et pénalisation de concentration variable (DVR). Il capture des comportements indétectables par les seuls indicateurs de fidélité ou de stabilité, notamment lorsque Random Forest et LightGBM sur Elliptic atteignent un F1 comparable tout en produisant des DVR très différents (0,65 contre 0,82). La principale limite, discutée en section 7.5, est la nécessité d'une curation experte du catalogue.

RQ3 (utilité par convergence d'agents LLM). La convergence décisionnelle, mesurée par le κ inter-agents et le κ LLM-versus-baseline-ML, constitue un substitut scalable d'une étude utilisateur humaine. Le passage de $\kappa \approx 0$ en C1 à $\kappa \approx 1{,}00$ en C2 ne se réduit pas à un effet de *prompt enrichment* (la condition C1 contient déjà les données brutes complètes), mais mesure bien l'apport informatif spécifique de l'explication.

7.2 H1, H2, H3

H1 (orthogonalité des axes). L'hypothèse est confirmée. GraphLIME sur GraphSAGE/Elliptic remporte certaines règles BRAS tout en restant inférieur à IG en fidélité. LIME sur Ethereum/Random Forest présente une Comprehensiveness honorable (0,394) mais un Lipschitz normalisé catastrophique (0,095). SHAP/LightGBM sur Elliptic atteint un Comprehensiveness élevé tout en produisant un DVR de 0,82. Ces dissociations valident l'investissement dans un cadre à quatre axes plutôt qu'un score scalaire unique.

H2 (paradoxe SMOTE confirmé). F1 et AUC sont préservés par SMOTE, mais Comprehensiveness chute de 12 à 31% et BRAS jusqu'à 18%. Le mécanisme, conforme à Zafar & Wu (2026), est un *covariate shift* induit par les exemples synthétiques. En interpolant linéairement dans l'espace des attributs, SMOTE crée des instances qui violent les corrélations naturelles du domaine, et le classifieur apprend une frontière s'appuyant sur ces régions interpolées. Les attributs jugés *importants* ne correspondent plus à ceux dont la suppression dégrade réellement la prédiction (chute de Comp), ni aux règles métier réelles (chute de BRAS). H2 est étendue empiriquement aux données blockchain (Bitcoin et Ethereum). Implication pratique : privilégier le *cost-sensitive learning* au sur-échantillonnage synthétique.

H3 (effet des explications sur les LLM). La chaîne ordinale $\kappa_{C1} < \kappa_{C2} < \kappa_{C3}$ est partiellement vérifiée. La première inégalité est massivement confirmée (transition de ~ 0 à 1,00). La seconde s'inverse parfois (Elliptic high-BRAS : $\kappa_{C3} = 0{,}83 < \kappa_{C2} = 1{,}00$). Cette inversion n'infirme pas H3, elle en raffine la lecture : l'indice d'incertitude calibré ajouté en C3 Guo *et al.* (2017) rouvre légitimement la dissidence sur les instances ambiguës. Une variance contrôlée du κ en C3 est précisément l'effet attendu d'un indicateur de confiance bien calibré.

7.3 INSIGHT CENTRAL : TRANSITION $\kappa_{\text{ML vs LLM}} : 0 \rightarrow 1$

Le résultat le plus important du mémoire est la transition $\kappa_{\text{ML vs LLM}} : 0 \rightarrow 1$ entre C1 et C2 (figure 6.8). Cette transition suggère une définition opérationnelle nouvelle : *une bonne explication est une explication qui fait converger un panel d'agents LLM vers les décisions du classifieur de référence*. Cette définition est scalable (quelques dollars par dizaine d'agents), reproductible (graines fixées) et alignée sur la pratique de conformité (lire l'explication et décider). Le $\kappa_{\text{ML vs LLM}}$ en C2 devient alors un score d'utilité directement comparable entre méthodes XAI.

Le corollaire est économique. Si le LLM *intègre* l'explication plutôt qu'il ne *raisonne* sur les données brutes, un modèle modeste suffit dès lors que l'explication est de qualité. Nos résultats sont compatibles avec cette lecture : en C2, les agents convergent presque uniformément quel que soit le modèle. En C1, où les agents doivent raisonner à partir des données brutes, des écarts inter-agents subsistent. Conséquence opérationnelle directe : le coût marginal d'un LLM premium ne se justifie pas pour un pipeline de conformité bien instrumenté en XAI, et la pile *classifieur* + *explainer* + *LLM* devient une pile dont l'*explainer* est le maillon critique où l'investissement marginal doit se concentrer.

7.4 POSITIONNEMENT PAR RAPPORT AUX TRAVAUX EXISTANTS

Quantus et OpenXAI Hedström *et al.* (2023); Agarwal *et al.* (2022) partagent l'esprit de nos modules 1 et 2 (métriques formelles standardisées), mais ne couvrent pas le domaine blockchain, n'incluent pas d'axe d'alignement domaine et n'instrumentent pas de validation utilisateur scalable par LLM. xFraud et SEFraud Rao *et al.* (2021); Li *et al.* (2024) reposent sur des modèles *self-explainable* ; le cadre BlockXAI-Eval est neutre vis-à-vis de l'origine de l'explication, ce qui en fait un outil de comparaison utilisable pour arbitrer entre post-hoc

et self-explainable. L'étude SMOTE (section 6.6) étend explicitement le cadre conceptuel de Zafar & Wu (2026) aux données blockchain, confirmant la généricité du paradoxe à travers les domaines.

7.5 IMPLICATIONS PRATIQUES

Le cadre fournit un pipeline reproductible auditable avant déploiement (CSV de scores, heatmaps, table de gagnants directement intégrables dans un dossier d'audit), un score BRAS aligné sur les exigences émergentes de gouvernance IA (AI Act, BCBS), et un benchmark partageable destiné à la communauté de recherche. Quatre limites pratiques restent à expliciter avant déploiement opérationnel : (i) le catalogue BRAS doit être curé par chaîne, le travail sur Bitcoin ne se transposant pas directement à Ethereum (DVR nul rapporté honnêtement) ; (ii) l'effectif $N = 5$ du Module 4 appelle une montée en échelle ; (iii) l'instabilité extrême de LIME sur Ethereum (Lipschitz brut $> 10^5$) interdit son déploiement seul sur ce type de données ; (iv) l'anomalie GNNExplainer/T-GCN (Comprehensiveness négatif) appelle une investigation complémentaire. Ces limites sont organisées dans la section 7.6 ci-dessous.

7.6 MENACES À LA VALIDITÉ ET LIMITATIONS

Nous adoptons la typologie classique de Wohlin et coll. (validité de construction, interne, externe, conclusion), enrichie d'une sous-section finale consacrée aux limites assumées. L'objectif est de prévenir toute surinterprétation des conclusions du chapitre 6 et de la discussion précédente, et de préparer les pistes de travaux futurs du chapitre 8.

7.6.1 VALIDITÉ DE CONSTRUCTION

Notre cadre ramène la qualité d'une explication à quatre axes (fidélité, stabilité, alignement métier, utilité humaine). Ce choix laisse délibérément de côté l'évaluation *counterfactual*, l'attribution causale au sens de Pearl et la *comprehensibility* subjective mesurée par questionnaires, dont l'opérationnalisation aurait excédé le périmètre d'un mémoire de maîtrise. Le module de fidélité repose sur un *masking* par médiane d'entraînement, choix qui se justifie pour les variables standardisées d'Elliptic (médiane ≈ 0) et pour les distributions asymétriques d'Ethereum (où l'imputation par zéro produirait des instances hors-distribution). Une étude de sensibilité comparant zéro, moyenne et médiane reste une extension pertinente. Le catalogue BRAS (quatorze règles sur Ethereum, ensemble restreint sur Elliptic en raison de l'anonymisation) reflète la sensibilité du chercheur ; nous l'avons rendu inspectable (annexe B) et une validation par accord interjuges demeure souhaitable. Par construction, BRAS est relatif à un catalogue explicité, et non une mesure absolue de plausibilité. Enfin, la substitution d'un panel d'agents LLM à un utilisateur humain est la simplification la plus discutable du cadre (voir section 7.6.2).

7.6.2 VALIDITÉ INTERNE

Le protocole repose sur un unique tirage aléatoire (graine 42 pour tous les composants stochastiques), ce qui garantit la reproductibilité bit-à-bit mais confond la variance d'estimation avec l'effet de condition. Idéalement, dix à trente graines distinctes auraient permis un intervalle de confiance bootstrap. Un *bootstrap interne* sur les attributions (CoV en stabilité) capture une partie de cette variance, mais pas l'effet conjoint de la séparation train/test et de l'initialisation des modèles. Plusieurs hyperparamètres ont été fixés par convention : $\sigma = 0{,}1$ pour Lipschitz et Infidelity (Yeh et coll., 2019), qui devient quasiment imperceptible sur les variables Ethereum couvrant plusieurs ordres de grandeur, expli-

quant les valeurs extrêmes pour LIME ; $k = 5$ pour Comp@k et Suff@k, dont le balayage $k \in \{1, 3, 5, 10\}$ (annexe D) montre que la hiérarchie des explicateurs reste stable. L'absence d'étude utilisateur humaine est la limite la plus saillante : les agents LLM, à $N = 5$ par condition, sont un proxy directionnel et non un substitut d'analyste expert. Le saut $\kappa : 0 \to 1$ entre C1 et C2 mesure un effet directionnel utile sans calibrer la magnitude pour un humain réel. En particulier, l'affirmation d'une quasi-indiscernabilité entre les agents LLM et la baseline ML doit être lue avec prudence : à $N = 5$ par condition, aucun test inférentiel formel ne peut l'établir, et nous la présentons comme une tendance forte constatée sur l'échantillon expérimental, non comme une équivalence statistiquement démontrée. Enfin, les modèles GNN n'ont été évalués que sur Elliptic (Ethereum est tabulaire, sans structure de graphe explicite), ce qui empêche d'isoler complètement les effets de typologie de modèle des effets de typologie de chaîne.

7.6.3 VALIDITÉ EXTERNE

L'étude porte sur deux datasets représentant deux modèles transactionnels distincts (Bitcoin UTXO, Ethereum compte). Ce périmètre est conceptuellement diversifié mais demeure modeste relativement à l'écosystème blockchain contemporain : Solana, les chaînes de couche 2 (Arbitrum, zkSync), Avalanche, n'ont pas été incluses. Ces écosystèmes présentent des particularités (granularité temporelle, coût marginal d'une interaction, disponibilité d'étiquettes) susceptibles de modifier les conclusions, et leur transposition demanderait une reconstruction des jeux de données et une révision du catalogue BRAS. Une seconde limite concerne la variance entre datasets : LIME présente une stabilité catastrophique sur Ethereum (Lipschitz $\geq 10^5$) mais acceptable sur Elliptic ($\leq 0{,}1$), différence principalement due à la standardisation Elliptic. Nous publions une version log-normalisée bornée pour atténuer ce biais, sans supprimer le phénomène sous-jacent : la sensibilité d'un

explicateur dépend de la nature des caractéristiques, et il convient de présenter les résultats par dataset. Enfin, la fraude blockchain est plurielle (mélangeurs, *rug-pulls*, *pump-and-dump*, hameçonnage, exploits de contrats) ; nos conclusions sont indexées sur les typologies couvertes par Elliptic et Ethereum Fraud, et une typologie différente (par exemple, attaques de gouvernance DeFi) pourrait inverser certaines préférences relatives.

7.6.4 VALIDITÉ DE CONCLUSION

Avec $N = 5$ instances par condition au Module 4 (et $N \in \{50, 100\}$ pour les Modules 1 à 3), l'inférence statistique formelle est fortement contrainte. Pour les Modules 1, 2 et 3, un test de Wilcoxon ou un bootstrap par rééchantillonnage aurait été envisageable ; nous avons choisi une lecture descriptive car les écarts (BRAS 0,874 pour T-GCN+IG contre 0,503 pour LGB+SHAP) sont d'une magnitude qui rend la significativité formelle peu informative. Une publication ultérieure intégrera ces tests pour les comparaisons à faible écart. Au Module 4, $N = 5$ rend toute inférence formelle inappropriée : les écarts inter-conditions, notamment $\kappa : 0 \to 1$ entre C1 et C2, sont visuellement robustes mais ne sauraient être confirmés par un test inférentiel. Nous qualifions donc ces conclusions de *constatées sur l'échantillon expérimental* et non de *généralisées en distribution*. L'ECE rapporté à $N = 5$ est mathématiquement mal défini (binning à grain grossier) et présenté à titre d'ordre de grandeur seulement.

7.6.5 LIMITES ASSUMÉES

Cinq limites résultent de choix de périmètre et demeurent assumées. *Premièrement*, BRAS dépend du catalogue : cette dépendance est intrinsèque (mesure de l'alignement à des *a priori* explicités, non d'une vérité de terrain), et le catalogue versionné est conçu

pour être discuté et révisé. Élaboré par un seul chercheur, ce catalogue gagnerait à être validé ultérieurement par plusieurs experts du domaine (analystes de conformité, enquêteurs anti-blanchiment), ce qui renforcerait sa robustesse et réduirait le biais de construction associé. *Deuxièmement*, le Module 4 repose sur des appels facturés à OpenRouter vers Claude Opus 4.7, Gemini 3.1 Pro et GPT 5.4. Cette dépendance commerciale soulève un enjeu de reproductibilité à moyen terme (mises à jour silencieuses) ; nous archivons les réponses brutes en JSON (annexe C) et publions les hash de version. Coût total observé de l'évaluation : 44,70$ USD pour 4,21 M jetons. *Troisièmement*, le cadre ne traite pas la robustesse adversariale (attaque de scaffolding de Slack et coll., 2020), extension naturelle identifiée au chapitre 8. *Quatrièmement*, l'évaluation en latence n'a pas été réalisée : les temps cumulés sont rapportés (annexe E) mais la latence en condition d'inférence unique reste à mesurer, en particulier pour LIME. *Cinquièmement*, l'article QRS 2026, désormais accepté et présenté, couvre le même socle empirique ; le mémoire ne constitue pas une double soumission (l'article cible un format conférence resserré de huit pages, le mémoire déploie le détail expérimental et les annexes complètes), avec autoréférence explicite (Oumbe Fokou, Salhab et Jaafar, 2026).

CHAPITRE VIII
CONCLUSION ET TRAVAUX FUTURS

8.1 SYNTHÈSE DES CONTRIBUTIONS

Ce mémoire répond à une question simple : comment évaluer rigoureusement la qualité d'une explication XAI dans le contexte de la détection de fraude sur blockchains publiques, où le déséquilibre de classes, la diversité des architectures (tabulaires et graphes temporels) et l'absence d'utilisateur expert accessible compliquent l'évaluation ? Quatre contributions structurent notre réponse. *Premièrement*, un **cadre multi-dimensionnel** articulé en quatre axes (fidélité, stabilité, alignement métier, utilité humaine) opérationnalisés par sept métriques, produisant un profil de qualité plutôt qu'un score unique. *Deuxièmement*, le score **BRAS** combinant alignement aux règles métier (RAS) et pertinence des variables dominantes (DVR), qui reformule l'alignement métier comme un calcul reproductible à partir d'un catalogue explicité. *Troisièmement*, un **protocole d'évaluation d'utilité fondé sur des agents LLM**, dans sa version 3 intégrant trois conditions (C1 prédiction seule, C2 prédiction et explication, C3 prédiction, explication et bande d'incertitude). *Quatrièmement*, une **étude empirique extensive** sur quatorze configurations (modèle, explicateur, dataset), couvrant Bitcoin (Elliptic) et Ethereum (Ethereum Fraud Detection), cinq explicateurs (SHAP, LIME, IG, GNNExplainer, GraphLIME), avec comparaison à une baseline ML linéaire. L'ensemble du code, des données prétraitées, des résultats CSV et des notebooks est publié sur GitHub (`xai-eval-blockchain-fraud`), avec hash de versions LLM et réponses brutes JSON.

8.2 RÉPONSES AUX QUESTIONS DE RECHERCHE

RQ1 appelle quatre dimensions interopérables, dont la nécessité est démontrée par des dissociations empiriques (LIME apparaît acceptable en fidélité mais sa stabilité catastrophique sur Ethereum la disqualifie pour un déploiement opérationnel ; information masquée par un cadre uni-axial), et dont la suffisance couvre les desiderata les plus invoqués dans la littérature sans dégénérescence statistique entre axes. L'omission d'un axe causal explicite reste reconnue comme extension future. **RQ2** est portée par BRAS, dont les résultats montrent une hiérarchie informative (GNN+IG sur Elliptic à 0,87, LGB+SHAP retombant à 0,50 malgré une fidélité élevée, en raison d'un DVR de 0,82) ; la dépendance au catalogue est une propriété revendiquée et non un défaut. **RQ3** appelle une réponse nuancée : les agents LLM fournissent un signal directionnel cohérent (Decision Accuracy passant de $\approx 0{,}60$ à $\approx 0{,}94$ sur Ethereum en C2, κ d'accord avec la baseline ML sautant de ≈ 0 à ≈ 1), mais leur caractère trop déférent les qualifie comme proxy de *pré-évaluation* plutôt que substitut d'analyste humain expert. Le protocole révèle qu'une explication a un effet, sans en calibrer la magnitude pour un humain réel.

8.3 PERSPECTIVES DE RECHERCHE

Passage à $N \geq 100$ et étude humaine. Porter le Module 4 à au moins 100 instances par condition restaurerait la puissance statistique pour des tests inférentiels formels (Wilcoxon apparié C1 contre C2) et un calcul rigoureux de l'ECE avec binning adéquat. À budget contraint, deux stratégies sont envisageables : restreindre le panel d'agents à un seul (GPT 5.4 suffit pour le signal directionnel), ou pratiquer une stratification active sur les instances proches de la frontière. En parallèle, une étude impliquant cinq à dix analystes de conformité (partenariat institutionnel ou prestataire de *blockchain analytics*) confronterait

décisions et délais humains aux sorties LLM, validant ou caractérisant l'écart entre proxy et expert.

Extension à Layer-2 et Solana. Les chaînes de couche 2 (Polygon, Arbitrum, Optimism) et Solana présentent des typologies d'attaque distinctes (bots *sandwich*, exploits de *liquidity pools*, *just-in-time liquidity*). La transposition demanderait une reconstitution de datasets étiquetés (peu d'équivalents publics à Elliptic), une adaptation des architectures aux graphes très denses, et une extension du catalogue BRAS. Le caractère non automatique de la généralisation est déjà visible dans le contraste LIME entre Elliptic et Ethereum.

Intégration en pipeline de conformité production. Trois enjeux concrets : une *suite de régression* automatique à chaque réentraînement, un *seuillage des alertes* (BRAS sous 0,70 ou Lipschitz dépassant un majorant déclenchant une alerte MLOps), et l'*exposition en service* via API REST permettant à d'autres équipes d'interroger la qualité d'une explication sans accès au modèle sous-jacent. Un travail sur la latence (hors périmètre du présent mémoire) conditionne la mise en production.

Évaluation adversariale. Slack et coll. (2020) ont montré que SHAP et LIME peuvent être attaqués par scaffolding. Un cinquième axe *adversarial robustness*, opérationnalisé par un *stress test* entraînant un méta-modèle scaffold et mesurant la dégradation conjointe de BRAS, Comprehensiveness et Lipschitz, fournirait une dimension préventive aujourd'hui absente.

Métrique d'explicabilité causale. Notre cadre mesure des contributions corrélatives, sans distinguer corrélation et causalité. Un sixième axe causal au sens de Pearl (*do*-calculus),

incorporant les décompositions causales des Shapley values (Janzing et coll., 2020 ; Heskes et coll., 2020 ; Frye et coll., 2021), comblerait une lacune théorique reconnue et renforcerait la défendabilité juridique des explications produites.

NOTE FINALE

Ce mémoire propose un cadre, un outil et une étude empirique qui contribuent à déplacer le problème de l'explicabilité en détection de fraude blockchain d'un terrain principalement qualitatif vers un terrain partiellement instrumenté, où des comparaisons reproductibles deviennent possibles, sans prétendre clore une question dont les limites assumées à la section 7.6 bornent honnêtement la portée.

BIBLIOGRAPHIE

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. & Kim, B. (2018). Sanity checks for saliency maps. Dans *Advances in Neural Information Processing Systems*, Vol. 31.

Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M. & Lakkaraju, H. (2022). OpenXAI : Towards a transparent evaluation of model explanations. Dans *Advances in Neural Information Processing Systems*, Vol. 35.

Alarab, I., Prakoonwit, S. & Nacer, M. I. (2020). Competence of graph convolutional networks for anti-money laundering in Bitcoin blockchain. Dans *Proc. 5th Int. Conf. Machine Learning Technologies*, pp. 23–27.

Alvarez-Melis, D. & Jaakkola, T. S. (2018). On the robustness of interpretability methods. Dans *ICML Workshop on Human Interpretability in Machine Learning*.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J. *et al.* (2020). Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Bhatt, U., Xiang, A., Sharma, S., Weller, A. *et al.* (2020). Explainable machine learning in deployment. Dans *Proc. ACM Conf. Fairness, Accountability, and Transparency*, pp. 648–657.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brown, T. B., Mann, B., Ryder, N. *et al.* (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chainalysis (2024). *The 2024 Crypto Crime Report*. Chainalysis Inc. Repéré le 2026-04-20, à https://www.chainalysis.com/reports/

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE : Synthetic minority over-sampling technique. *J. Artificial Intelligence Research*, *16*, 321–357.

Chen, L., Peng, J., Liu, Y., Li, J., Xie, F. & Zheng, Z. (2020). Phishing scams detection in Ethereum transaction network. *ACM Trans. Internet Technology*, *21*(1), 1–16.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L. & Juels, A. (2020). Flash boys 2.0 : Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. Dans *IEEE Symp. Security and Privacy*, pp. 910–927.

Demertzis, K., Iliadis, L., Tziritas, N. & Kikiras, P. (2021). Explainable artificial intelligence for cyber threat intelligence in blockchain forensics. *Neural Computing and Applications*, *33*, 13647–13665.

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv :1702.08608*.

Elkan, C. (2001). The foundations of cost-sensitive learning. Dans *Proc. 17th Int. Joint Conf. Artificial Intelligence*, pp. 973–978.

European Parliament and Council (2023). *Regulation (EU) 2023/1114 on Markets in Crypto-Assets (MiCA)*. Official Journal of the European Union.

Farrugia, S., Ellul, J. & Azzopardi, G. (2020). Detection of illicit accounts over the Ethereum blockchain. *Expert Systems with Applications*, *150*, 113318.

Financial Action Task Force (2021). *Updated Guidance for a Risk-Based Approach to Virtual Assets and Virtual Asset Service Providers*. FATF.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. (2017). Neural message passing for quantum chemistry. Dans *Proc. 34th Int. Conf. Machine Learning*, pp. 1263–1272.

Goyal, R., Manoov, R. & Karunakaran, S. (2022). Detecting fraudulent accounts on Ethereum : A supervised approach. *Journal of Information Security and Applications*,

68, 103231.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42.

Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. (2017). On calibration of modern neural networks. Dans *Proc. 34th Int. Conf. Machine Learning*, pp. 1321–1330.

Hamilton, W. L., Ying, R. & Leskovec, J. (2017). Inductive representation learning on large graphs. Dans *Advances in Neural Information Processing Systems*, Vol. 30.

Hedström, A., Weber, L., Krakowczyk, D. *et al.* (2023). Quantus : An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *J. Machine Learning Research*, *24*(34), 1–11.

Huang, Q., Yamada, M., Tian, Y., Singh, D. & Chang, Y. (2022). GraphLIME : Local interpretable model explanations for graph neural networks. *IEEE Trans. Knowledge and Data Engineering*, *35*(7), 6968–6972.

Kadavath, S., Conerly, T., Askell, A. *et al.* (2022). Language models (mostly) know what they know. *arXiv preprint arXiv :2207.05221*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. (2017). LightGBM : A highly efficient gradient boosting decision tree. Dans *Advances in Neural Information Processing Systems*, Vol. 30.

Kipf, T. N. & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. Dans *Int. Conf. Learning Representations*.

Klaise, J., Van Looveren, A., Vacanti, G. & Coca, A. (2021). Alibi Explain : Algorithms for explaining machine learning models. *J. Machine Learning Research*, *22*(181), 1–7.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*,

35.

Kokhlikyan, N., Miglani, V., Martin, M. *et al.* (2020). Captum : A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv :2009.07896*.

Li, K., Yang, Y. *et al.* (2024). SEFraud : Graph-based self-explainable fraud detection via interpretable mask learning. *Proc. ACM SIGKDD*.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. & Zhu, C. (2023). G-Eval : NLG evaluation using GPT-4 with better human alignment. Dans *Proc. Conf. Empirical Methods in Natural Language Processing*, pp. 2511–2522.

Liu, Z., Dou, Y., Yu, P. S., Deng, Y. & Peng, H. (2021). Explainable graph-based fraud detection via neural meta-graph search. Dans *Proc. 30th ACM Int. Conf. Information and Knowledge Management*, pp. 4414–4418.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Dans *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4765–4774.

Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H. & Zhang, X. (2020). Parameterized explainer for graph neural network. Dans *Advances in Neural Information Processing Systems*, Vol. 33.

Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G. M. & Savage, S. (2013). A fistful of Bitcoins : Characterizing payments among men with no names. Dans *Proc. ACM Internet Measurement Conf.*, pp. 127–140.

Möser, M. & Narayanan, A. (2022). Resurrecting address clustering in Bitcoin. *Financial Cryptography and Data Security*.

Naiseh, M., Al-Thani, D., Jiang, N. & Ali, R. (2023). How the different explanation classes impact trust calibration : The case of clinical decision support systems. *Int. J.*

Human-Computer Studies, *169*, 102941.

Nauta, M., Trienes, J., Pathak, S. *et al.* (2023). From anecdotal evidence to quantitative evaluation methods : A systematic review on evaluating explainable AI. *ACM Computing Surveys*, *55*(13s), 1–42.

Pareja, A., Domeniconi, G., Chen, J., Ma, T., Suzumura, T., Kanezashi, H., Kaler, T., Schardl, T. B. & Leiserson, C. E. (2020). EvolveGCN : Evolving graph convolutional networks for dynamic graphs. Dans *Proc. AAAI Conf. Artificial Intelligence*, Vol. 34, pp. 5363–5370.

Pocher, N., Zichichi, M., Merlini, F. & Bonomi, S. (2023). Detecting anomalous cryptocurrency transactions : An AML/CFT application of machine learning-based forensics. *Electronic Markets*, *33*(1), 37.

Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P. & Petitcolas, F. (2021). Explainable machine learning for fraud detection. *IEEE Computer*, *54*(10), 49–59.

Rao, S. X., Zhang, S., Han, Z., Zhang, Z., Min, W., Cheng, Z., Shan, Y., Zhao, Y. & Zhang, C. (2021). xFraud : Explainable fraud transaction detection. *Proc. VLDB Endowment*, *15*(3), 427–436.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). "Why should I trust you ?" : Explaining the predictions of any classifier. Dans *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 1135–1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Slack, D., Hilgard, S., Jia, E., Singh, S. & Lakkaraju, H. (2020). Fooling LIME and SHAP : Adversarial attacks on post hoc explanation methods. Dans *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, pp. 180–186.

Sundararajan, M., Taly, A. & Yan, Q. (2017). Axiomatic attribution for deep networks. Dans *Proc. 34th Int. Conf. Machine Learning*, pp. 3319–3328.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. & Bengio, Y. (2018). Graph attention networks. Dans *Int. Conf. Learning Representations*.

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T. & Leiserson, C. E. (2019). Anti-money laundering in Bitcoin : Experimenting with graph convolutional networks for financial forensics. Dans *KDD Workshop on Anomaly Detection in Finance*.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I. & Ravikumar, P. K. (2019). On the (in)fidelity and sensitivity of explanations. Dans *Advances in Neural Information Processing Systems*, Vol. 32.

Ying, Z., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. (2019). GNNExplainer : Generating explanations for graph neural networks. Dans *Advances in Neural Information Processing Systems*, Vol. 32.

Yuan, H., Yu, H., Gui, S. & Ji, S. (2022). Explainability in graph neural networks : A taxonomic survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *45*(5), 5782–5799.

Yuan, H., Yu, H., Wang, J., Li, K. & Ji, S. (2021). On explainability of graph neural networks via subgraph explorations. Dans *Proc. 38th Int. Conf. Machine Learning*, pp. 12241–12252.

Zafar, M. B. & Wu, Y. (2026). Methodological challenges of explainable AI in fraud detection : A systematic review. *ACM Computing Surveys*. In press.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Xing, E. P., Gonzalez, J. E. & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. Dans *Advances in Neural Information Processing Systems*, Vol. 36.

Zhou, J., Gandomi, A. H., Chen, F. & Holzinger, A. (2021). Evaluating the quality of machine learning explanations : A survey on methods and metrics. *Electronics*, *10*(5), 593.

APPENDICE A
CARACTÉRISTIQUES DÉTAILLÉES DES DATASETS

Cette annexe complète le chapitre 5 en exposant la généalogie et la structure des deux jeux de données mobilisés. Elle vise à permettre une reproduction intégrale et à expliciter le contraste structurel qui motive leur sélection conjointe.

A.1 DATASET ELLIPTIC BITCOIN

Origine. Le dataset Elliptic Bitcoin a été publié en 2019 par Elliptic Co. en collaboration avec des chercheurs du MIT et d'IBM Research (Weber *et al.*, 2019). Il constitue le plus grand graphe transactionnel public étiqueté de Bitcoin, couvrant environ deux semaines de transactions échantillonnées entre 2014 et 2017. Les transactions sont les nœuds d'un graphe orienté, les arêtes représentant le flux UTXO. Les étiquettes proviennent d'investigations OSINT et de signatures comportementales validées par les analystes Elliptic.

Schéma des features. Chaque transaction est décrite par 166 caractéristiques numériques : les 94 premières (`tx_feat_0` à `tx_feat_93`) sont *locales* (volume, frais, entrées/sorties, intervalles temporels), anonymisées par transformation affine non publiée ; les 72 suivantes (`tx_feat_94` à `tx_feat_165`) sont *agrégées* sur le voisinage à un saut (moyenne, écart-type, quantiles). L'ensemble est standardisé (moyenne nulle, variance unitaire), ce qui borne la sensibilité des explicateurs locaux.

Graphe et splits. 203 769 nœuds, 234 355 arêtes, 49 *timesteps* temporels. Étiquetage : 4 545 illicites (2,2%), 42 019 licites (20,6%), 157 205 inconnus. L'apprentissage est restreint

aux 46 564 nœuds étiquetés. Le *dark market shutdown* au timestep 43 provoque un *concept drift* bien documenté (Alarab *et al.*, 2020). Partition temporelle stricte : timesteps 1 à 34 (entraînement, 29 894 nœuds), 35 à 39 (validation), 40 à 49 (test, 16 670 nœuds).

A.2 DATASET ETHEREUM FRAUD

Origine. Le dataset a été constitué par Farrugia *et al.* (2020) à partir de 9 841 comptes Ethereum, dont 2 179 frauduleux (22,1%) et 7 662 légitimes, étiquetés via listes Etherscan/MyCrypto, rapports de hameçonnage et investigations manuelles. Représentation tabulaire par compte (pas de structure de graphe fournie).

Schéma des 45 features nommées. Contrairement à Elliptic, les variables sont nommées et interprétables, ce qui autorise un catalogue BRAS sémantiquement défendable. Nous les regroupons en six familles synthétisées dans la table A.1.

TABLEAU A.1 : Six familles de features Ethereum (indices entre parenthèses).

Famille	Features (extraits)
Temporelles (0–2)	Avg min between sent/received tnx ; Time Diff first-last
Volume (3, 4, 17–19, 39)	Sent/Received tnx ; Total Ether sent/received ; Min value received
Valeur Ether (8–13)	Min/max/avg value sent ; min/max/avg value received
Contrats (5, 14–16, 20, 44)	Created contracts ; value sent to contracts ; total ether balance
Diversité (6, 7, 25, 26)	Unique sent/received addresses ; ERC20 unique sent/rec addr
ERC-20 (21–38)	18 caractéristiques sur les transactions de jetons ERC-20
Ratios (40–44)	Total ERC20 tnxs ; most sent/received token type ; total balance

Split. En l'absence d'ordre chronologique exploitable, nous adoptons un split stratifié aléatoire (graine 42), 70/15/15, préservant la proportion 22,1% d'illicites dans chaque sous-ensemble.

A.3 JUSTIFICATION DU CHOIX CONJOINT

Quatre contrastes motivent la sélection conjointe : modèles transactionnels distincts (UTXO contre compte), représentations distinctes (graphe contre tabulaire, autorisant l'évaluation GNN sur Elliptic), interprétabilité des variables (anonymisée contre nommée, ce dernier point autorisant un catalogue BRAS sémantique), et amplitude de déséquilibre (2,2% contre 22,1%). Ces contrastes permettent d'isoler partiellement les effets de chaîne, de modèle et de régime de classes.

APPENDICE B

PSEUDO-CODE DES MÉTRIQUES

Cette annexe formalise les trois métriques centrales du cadre : Comprehensiveness@k (fidélité), Lipschitz local (stabilité) et BRAS (alignement métier). Les autres métriques (Infidelity, Sufficiency, log-normalisation, pipeline LLM) sont décrites textuellement au chapitre 4.

B.1 COMPREHENSIVENESS@K

Mesure la chute de probabilité lorsque les k features les plus saillantes sont retirées par imputation à la médiane (Yeh et coll., 2019).

Entrées : Modèle f, explicateur Φ, instance $x \in \mathbb{R}^d$, entier k, médiane $m \in \mathbb{R}^d$

Output : $\Delta_{\text{comp}} \in \mathbb{R}$

$p_0 \leftarrow f(x)[c]$; `// c : classe prédite`

$A \leftarrow \Phi(f, x)$;

$\mathscr{I}_k \leftarrow$ argtop-$k(|A|)$;

$x' \leftarrow x$;

pour chaque $j \in \mathscr{I}_k$ **faire**

  $x'_j \leftarrow m_j$;

fin

$p_1 \leftarrow f(x')[c]$;

retourner $p_0 - p_1$;

Algorithme B.1 : Calcul de Comprehensiveness@k

Complexité : $O(d + C_f)$.

B.2 CONSTANTE DE LIPSCHITZ LOCALE

Borne inférieure empirique par échantillonnage k-NN et perturbations gaussiennes ; mesure la sensibilité de Φ aux petites perturbations.

Entrées : Instance x, explicateur Φ, voisinage $\mathcal{V}_k(x)$, écart-type σ, perturbations K

Output : $L(x) \in \mathbb{R}_{\geq 0}$

$L \leftarrow 0$;

pour chaque $x' \in \mathcal{V}_k(x)$ **faire**

  pour $j \leftarrow 1$ **à** K **faire**

    $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$;

    $x'' \leftarrow x' + \varepsilon$;

    $L_j \leftarrow \|\Phi(x'') - \Phi(x')\|_2 / \max(\|x'' - x'\|_2, \texttt{eps})$;

    $L \leftarrow \max(L, L_j)$;

  fin

fin

retourner L;

Algorithme B.2 : Lipschitz local par k-NN

Complexité : $O(k \cdot K \cdot C_\Phi)$. Sur LIME, C_Φ inclut le réentraînement d'un modèle local par voisin, ce qui domine le budget.

B.3 BRAS (RAS + DVR)

Composite combinant couverture des règles métier (RAS) et pénalisation de la concentration sur une variable unique (DVR).

Entrées : $\{\Phi(x_i)\}_{i=1}^{N}$, catalogue $\mathscr{R} = \{r_1, \ldots, r_R\}$ (chaque r_j pointe vers $\mathscr{F}_j$), top-k, poids $\alpha = 0{,}5$

Output : $\mathrm{BRAS}, \mathrm{RAS}, \mathrm{DVR} \in [0,1]$

$n_{\mathrm{hit}} \leftarrow 0$;

pour $i \leftarrow 1$ **à** N **faire**

  $\mathscr{I}_k^{(i)} \leftarrow \text{argtop-}k(|\Phi(x_i)|)$;

  si $\exists\, j$: $\mathscr{I}_k^{(i)} \cap \mathscr{F}_j \neq \emptyset$ *et* r_j *satisfaite par* x_i **alors**

    $n_{\mathrm{hit}} \leftarrow n_{\mathrm{hit}} + 1$;

  fin

fin

$\mathrm{RAS} \leftarrow n_{\mathrm{hit}}/N$;

$\Phi_{\mathrm{moy}} \leftarrow (1/N)\sum_i |\Phi(x_i)|$;

$\mathrm{DVR} \leftarrow \max_j \Phi_{\mathrm{moy}}[j] / \sum_j \Phi_{\mathrm{moy}}[j]$;

$\mathrm{BRAS} \leftarrow \alpha \cdot \mathrm{RAS} + (1-\alpha)(1-\mathrm{DVR})$;

retourner $\mathrm{BRAS}, \mathrm{RAS}, \mathrm{DVR}$;

Algorithme B.3 : Calcul du score BRAS

Complexité : $O(N \cdot R \cdot k + N \cdot d)$. Le composite pénalise les explications concentrées sur une variable unique, reflétant l'*a priori* qu'une explication métier-pertinente mobilise plusieurs signaux complémentaires.

APPENDICE C

PROMPTS UTILISÉS POUR LES AGENTS LLM

Cette annexe reproduit les prompts du Module 4. Les trois conditions partagent une structure commune (message système d'analyste de conformité, message utilisateur dépendant de la condition, format JSON imposé). Les prompts sont en anglais (langue native d'entraînement des trois modèles cibles).

C.1 PROMPT C1 : PRÉDICTION SEULE

Message système.

```
You are a senior blockchain compliance analyst with
10+ years of experience in anti-money-laundering
(AML) investigations. You are reviewing a single
account/transaction record extracted from a public
blockchain dataset. Your role is to make a binary
classification decision: is this entity LICIT
(legitimate) or ILLICIT (fraudulent / suspicious /
criminally connected)? You must also report a
self-calibrated confidence in [0, 1]. Be precise,
conservative, and base your judgment strictly on
the information provided.
```

Message utilisateur.

```
Below is an entity from the {DATASET_NAME} dataset.
Decide whether it is LICIT or ILLICIT.

Feature vector (raw values):
{FEATURE_TABLE_AS_KEY_VALUE_PAIRS}

Respond strictly in the following JSON schema,
no extra text:
{
  "verdict":    "LICIT" | "ILLICIT",
  "confidence": <float in [0,1]>,
  "rationale":  "<short justification, <= 2 sentences>"
}
```

C.2 PROMPT C2 : PRÉDICTION ET EXPLICATION

Message système. Identique à C1.

Message utilisateur.

```
Below is an entity from the {DATASET_NAME} dataset.
A machine-learning fraud detector has produced a
prediction together with a feature-attribution
explanation (top-5 features by absolute contribution).
Decide whether the entity is LICIT or ILLICIT, taking
the explanation into account as a non-binding signal.
```

```
Feature vector (raw values):
{FEATURE_TABLE_AS_KEY_VALUE_PAIRS}

Model prediction:
- verdict:              {MODEL_VERDICT}
- probability_illicit: {MODEL_PROB}

Top-5 attributions (XAI method = {XAI_NAME}):
1. {feature_1}: weight = {w_1}
2. {feature_2}: weight = {w_2}
3. {feature_3}: weight = {w_3}
4. {feature_4}: weight = {w_4}
5. {feature_5}: weight = {w_5}
(positive weight => pushes toward ILLICIT)

Respond strictly in the following JSON schema:
{
  "verdict":       "LICIT" | "ILLICIT",
  "confidence":    <float in [0,1]>,
  "used_features": [<feature names relied upon>],
  "rationale":     "<short justification>"
}
```

C.3 PROMPT C3 : PRÉDICTION, EXPLICATION ET BANDE D'INCERTITUDE

Message système. Identique à C1, augmenté de :

```
When an uncertainty band is provided alongside the
model probability, treat it as an indicator of model
self-doubt. Reduce your own confidence proportionally
when the band is wide.
```

Message utilisateur.

```
Below is an entity from the {DATASET_NAME} dataset.
A machine-learning fraud detector has produced a
prediction with a confidence interval and a feature-
attribution explanation (top-5 features). Make a
binary decision and report a calibrated confidence
that integrates the model's uncertainty band.

Feature vector (raw values):
{FEATURE_TABLE_AS_KEY_VALUE_PAIRS}

Model prediction:
- verdict:              {MODEL_VERDICT}
- probability_illicit: {MODEL_PROB}
- uncertainty_band:    [{LOWER}, {UPPER}]  # 80% CI

Top-5 attributions (XAI method = {XAI_NAME}):
1. {feature_1}: weight = {w_1}
2. {feature_2}: weight = {w_2}
3. {feature_3}: weight = {w_3}
```

```
4. {feature_4}: weight = {w_4}
5. {feature_5}: weight = {w_5}

Respond strictly in the following JSON schema:
{
  "verdict":          "LICIT" | "ILLICIT",
  "confidence":       <float in [0,1]>,
  "used_features":    [<feature names relied upon>],
  "uncertainty_used": "<one sentence on band use>",
  "rationale":        "<short justification>"
}
```

C.4 PARAMÈTRES D'APPEL

Les trois agents sont interrogés via OpenRouter avec `temperature=0.2`, `top_p=0.9`, `max_tokens=300`, `seed=42` (OpenAI uniquement). Pour Claude et Gemini, trois tirages avec vote majoritaire (*self-consistency*). Identifiants exacts : `anthropic/claude-opus-4.7`, `google/gemini-3.1-pro-preview`, `openai/gpt-5.4`. Le parser applique trois stratégies de remédiation (extraction par regex, validation par JSON Schema Draft 7, relance unique corrective), avec un taux résiduel de `parse_error` inférieur à 1%. L'archive des réponses brutes est dans `experiments/results/module4v3_llm_raw_responses.csv`.

APPENDICE D

RÉSULTATS DÉTAILLÉS

Cette annexe regroupe les tables synthétiques des modules 1, 3 et 4. Les valeurs détaillées par instance et par graine, ainsi que les tables de stabilité (Module 2) et de comparaison ML/LLM intégrale (D.5 supprimée), sont disponibles dans les fichiers CSV du dépôt GitHub sous `experiments/results/`.

D.1 TABLE D.1 – MODULE 1 (FIDÉLITÉ, SYNTHÈSE À $k = 5$ ET $k = 10$)

Source : `module1_fidelity_results.csv`. Synthèse aux deux valeurs principales de k ; le balayage complet $k \in \{1, 3, 5, 10\}$ est dans le CSV. *Comp* et *Suff* : Comprehensiveness@k et Sufficiency@k. *Infid* : Infidelity Monte Carlo (indépendante de k).

TABLEAU D.1 : Module 1 – Fidélité, synthèse à $k \in \{5, 10\}$.

Modèle	Dataset	XAI	k	Comp	Suff	Infid
RF	Elliptic	SHAP	5	0,238	0,377	0,213
LGB	Elliptic	SHAP	5	0,439	0,427	0,436
RF	Ethereum	SHAP	5	0,149	0,294	0,228
LGB	Ethereum	SHAP	5	0,116	0,104	0,618
RF	Elliptic	LIME	5	0,341	0,697	0,415
LGB	Elliptic	LIME	5	0,322	0,846	0,637
RF	Ethereum	LIME	5	0,394	0,662	0,391
LGB	Ethereum	LIME	5	0,196	0,444	0,573
T-GCN	Elliptic	GNNExp	5	−0,035	0,114	0,128
T-GCN	Elliptic	IG	5	0,264	−0,113	1,505
T-GCN	Elliptic	GraphLIME	5	0,055	0,103	0,009
SAGE	Elliptic	GNNExp	5	0,229	0,045	0,102
SAGE	Elliptic	IG	5	0,280	−0,034	1,080
SAGE	Elliptic	GraphLIME	5	0,222	0,037	0,009

D.2 TABLE D.2 – MODULE 3 (BRAS)

Source : `module3_bras_summary.csv`. Tri par BRAS décroissant.

TABLEAU D.2 : Module 3 – BRAS détaillé, 14 configurations.

Modèle	Dataset	XAI	RAS	DVR	BRAS	N_{eval}
T-GCN	Elliptic	IG	0,928	0,18	0,874	50
SAGE	Elliptic	IG	0,920	0,20	0,860	50
SAGE	Elliptic	GraphLIME	0,928	0,24	0,844	50
T-GCN	Elliptic	GNNExp	0,944	0,26	0,842	50
RF	Elliptic	LIME	0,926	0,30	0,813	100
T-GCN	Elliptic	GraphLIME	0,912	0,32	0,796	50
SAGE	Elliptic	GNNExp	0,912	0,34	0,786	50
LGB	Ethereum	SHAP	0,518	0,00	0,759	100
RF	Ethereum	LIME	0,478	0,00	0,739	100
LGB	Elliptic	LIME	0,870	0,49	0,690	100
LGB	Ethereum	LIME	0,370	0,00	0,685	100
RF	Ethereum	SHAP	0,246	0,00	0,623	100
RF	Elliptic	SHAP	0,870	0,65	0,610	100
LGB	Elliptic	SHAP	0,826	0,82	0,503	100

D.3 TABLE D.3 – MODULE 4 V3 (LLM, AGRÉGÉ SUR LES AGENTS)

Source : `module4v3_llm_summary.csv`. DA moyenne et écart-type sur trois agents, ECE, EU (C2/C3 uniquement), κ d'accord avec le classifieur de référence.

TABLEAU D.3 : Module 4 v3 – LLM agrégé sur les trois agents.

Dataset	BRAS	Cond.	$\overline{\text{DA}}$	σ_{DA}	ECE	EU	κ
Elliptic	0,813 (haut)	C1	0,790	0,260	0,234	—	0,010
Elliptic	0,813 (haut)	C2	0,814	0,000	0,136	0,854	1,000
Elliptic	0,813 (haut)	C3	0,795	0,046	0,105	0,570	0,831
Elliptic	0,503 (bas)	C1	0,520	0,052	0,218	—	−0,040
Elliptic	0,503 (bas)	C2	0,830	0,017	0,119	0,835	0,960
Elliptic	0,503 (bas)	C3	0,843	0,005	0,105	0,546	0,986
Ethereum	0,759 (haut)	C1	0,602	0,211	0,164	—	0,072
Ethereum	0,759 (haut)	C2	0,940	0,000	0,045	0,601	1,000
Ethereum	0,759 (haut)	C3	0,940	0,000	0,118	0,462	1,000
Ethereum	0,623 (bas)	C1	0,617	0,215	0,140	—	0,067
Ethereum	0,623 (bas)	C2	0,926	0,016	0,050	0,544	0,959
Ethereum	0,623 (bas)	C3	0,933	0,006	0,102	0,419	0,986

Lecture transversale. Les tableaux D.1 à D.3 confirment trois constats : la fidélité est monotone en k pour Comp et Suff, à l'exception de GNNExplainer/T-GCN dont les valeurs négatives signalent une qualité médiocre ; BRAS discrimine fortement les configurations même à F1 comparable ; le saut de DA et de κ entre C1 et C2 est massif et reproductible, sa magnitude dépendant du niveau de BRAS sous-jacent. Les détails par instance et par graine, ainsi que les tables de stabilité (Module 2, 14 configurations avec Lipschitz brut/log/normalisé, τ de Kendall, CoV, identité) et de comparaison ML/LLM par classifieur, sont disponibles dans les fichiers CSV du dépôt GitHub.

APPENDICE E

CONFIGURATION MATÉRIELLE ET TEMPS DE CALCUL

E.1 MATÉRIEL

Station locale Windows 11 (Intel/AMD 8 cœurs / 16 logiques à $\approx$ 3,5 GHz, 32 Go RAM, GPU NVIDIA RTX 30/40 avec 8 à 12 Go VRAM, CUDA 12.*x*, Python 3.13) pour le développement, les modèles tabulaires, l'entraînement GNN et les modules 1 à 3. Deux VPS Contabo (Ubuntu Server 24.04 LTS, 4 vCPU, 16 Go RAM, sans GPU) dédiés au Module 4 pour bénéficier d'une connectivité plus régulière vers OpenRouter (latence médiane 35 ms contre 80 ms).

E.2 TEMPS DE CALCUL

TABLEAU E.1 : Temps de calcul des notebooks (temps mural).

Notebook	Durée	Environnement
00_setup_environment	$\approx$ 5 min	Local
01_eda_elliptic + 01b_eda_ethereum	$\approx$ 18 min	Local
02a_baselines_elliptic (RF, LGB, T-GCN, SAGE)	$\approx$ 25 min	Local (GPU)
02b_baselines_ethereum (RF, LGB)	$\approx$ 12 min	Local
03_xai_attributions (SHAP, LIME, IG, GNNExp, GraphLIME)	$\approx$ 40 min	Local
04_module1_fidelity	$\approx$ 45 min	Local
05_module2_stability (bootstrap $K = 200$)	$\approx$ 2 h 15	Local
06_module3_bras	$\approx$ 20 min	Local
07_exp_imbalance_smote_costsens	$\approx$ 30 min	Local
08_module4_llm_agents_v3 (3 agents, 3 conditions, $N = 5$)	$\approx$ 10 h 33	VPS Contabo
09_module4_ml_baseline (LR, RF)	$\approx$ 10 s	Local
Total cumulé	$\approx$ 16 h	—

Module 2 dominé par le bootstrap interne et le réentraînement local de LIME. Module 4 : $5 \times 3 \times 3 = 45$ requêtes par condition, $\times 3$ conditions $\times 2$ datasets $\times 2$ niveaux BRAS $= 540$ requêtes, plus $< 1\%$ de retries après parse.

E.3 COÛTS D'API

Coût total observé du Module 4 (cumul OpenRouter sur l'ensemble des exécutions) : \$44,70 USD pour environ 4,21 M jetons et 4 k requêtes. Répartition par fournisseur : Claude Opus 4.7 \$19,56 (1,56 M jetons, 1,23 k requêtes), Gemini 3.1 Pro Preview \$12,20 (1,56 M jetons, 1,22 k requêtes), GPT 5.4 \$9,12 (1,09 M jetons, 1,17 k requêtes). Projection linéaire à $N = 100$ par condition : $\approx$ \$900 USD pour la triade complète, ramené à \$185 USD si un seul agent est conservé (GPT 5.4 suffit pour le signal directionnel principal).

E.4 REPRODUCTIBILITÉ

Trois principes gouvernent la reproductibilité : gel des versions (toutes les dépendances épinglées dans `requirements.txt` du dépôt, voir liste complète sur GitHub), gel des graines (`seed=42` pour numpy, torch, random, sklearn, OpenAI ; stratégie *self-consistency* à trois tirages avec vote majoritaire pour Anthropic/Google qui n'exposent pas de seed déterministe), orchestration unifiée (`run_all.py` exécute séquentiellement les notebooks et vérifie les hash MD5 des CSV contre une table de référence à 10^{-6} près). Exécution complète $\approx$ 16 heures sur station équivalente, hors Module 4 parallélisable sur VPS. Publication sous licence MIT (`xai-eval-blockchain-fraud`), snapshot Zenodo archivé sous DOI 10.5281/zenodo.20210031.