AML-Bench-FR

Vers un référentiel d'évaluation des systèmes d'IA générative en matière de lutte contre le blanchiment de capitaux et le financement du terrorisme

AML-Bench-FR Project

Mai 2026 — Version 0.1

Abstract

Les systèmes d'IA générative augmentés par recherche (RAG) sont en cours de déploiement rapide dans les fonctions de conformité LCB-FT, sans qu'aucun cadre d'évaluation standardisé n'existe pour mesurer leur fiabilité dans ce domaine. Ce document propose la construction d'AML-Bench-FR, un benchmark de référence pour évaluer les capacités des LLM et des architectures RAG sur les tâches de conformité LCB-FT, avec un ancrage initial dans le droit français et européen (AMLD, AMLR, lignes directrices ACPR et TRACFIN, recommandations GAFI). Le benchmark adopte une approche multi-axes (sept capacités cognitives), une méthodologie de construction inspirée de GPQA et LegalBench (sourcing primaire, validation contradictoire experte, design adversarial), et un protocole d'évaluation holistique inspiré de HELM (au-delà de la seule accuracy : calibration, fidélité, robustesse, refus approprié). L'objectif est de produire un actif académique, technique et stratégique au service d'une régulation IA crédible dans le secteur de la compliance financière.

Contents

1. Contexte et motivation

1.1 Convergence réglementaire et technologique

Le paysage LCB-FT européen connaît une transformation structurelle. Le règlement européen AMLR, applicable au 10 juillet 2027, l'établissement de l'autorité européenne AMLA, et l'extension du périmètre des entités assujetties aux activités non-financières (agents immobiliers, professionnels du chiffre, marchands de biens précieux, prestataires de services sur actifs numériques) génèrent un besoin sans précédent en outils d'aide à la conformité. Simultanément, l'AI Act européen classe les systèmes d'IA déployés en compliance financière parmi les usages à haut risque, imposant des obligations d'évaluation, de documentation et de gestion des risques.

Cette double dynamique crée un appel d'air pour des solutions IA, mais aucune méthodologie de référence ne permet aujourd'hui à un acquéreur, un régulateur ou un auditeur d'évaluer objectivement les performances d'un système d'IA appliqué à la LCB-FT. Les benchmarks NLP généralistes (MMLU, BIG-bench) ne couvrent pas le domaine. Les benchmarks juridiques (LegalBench) restent ancrés en common law américain. Les benchmarks financiers (FinanceBench) couvrent l'analyse corporate, non la conformité réglementaire. Cette absence est un risque systémique : elle ouvre la porte à des déploiements d'IA non évalués sur des fonctions à fort enjeu juridique et opérationnel.

1.2 Positionnement

AML-Bench-FR vise à devenir le référentiel d'évaluation des systèmes d'IA — modèles de langage seuls et architectures RAG — sur les tâches de conformité LCB-FT en droit français et européen. Le projet poursuit trois objectifs simultanés :

Académique : produire un dataset et une méthodologie d'évaluation publiables, citables, reproductibles, contribuant à la littérature en NLP juridique et en évaluation des systèmes IA spécialisés.

Industriel : offrir aux éditeurs reg-tech, aux entités assujetties et à leurs auditeurs un outil de mesure objectif des performances et limites des systèmes IA en conformité.

Réglementaire : anticiper les attentes des autorités (ACPR, AMLA, autorités de marché) en matière d'évaluation d'IA à haut risque dans les fonctions de compliance, et fournir un cadre potentiellement adoptable comme référence supervisoire.

2. État de l'art

2.1 Benchmarks généralistes et spécialisés

MMLU et MMLU-Pro (Hendrycks et al., 2020 ; Wang et al., 2024) — couverture multi-domaine large, format QCM. Limites documentées : saturation par les modèles frontières, contamination probable des données d'entraînement, faible profondeur de raisonnement.

GPQA (Rein et al., 2023) — questions de niveau doctoral en sciences naturelles, conçues pour résister à la recherche web. Méthodologie de validation experte contradictoire à laquelle ce projet emprunte explicitement.

LegalBench (Guha et al., 2023) — 162 tâches juridiques, taxonomie IRAC (Issue, Rule, Application, Conclusion), construction collaborative par juristes. Référence structurelle pour AML-Bench-FR, mais ancrage common law et absence de couverture LCB-FT.

FinanceBench (Islam et al., 2023) — questions sur documents 10-K, approche evidence-based avec passages sources annotés. Inspiration pour la traçabilité documentaire mais hors scope LCB-FT.

HELM (Liang et al., 2022, Stanford CRFM) — framework d'évaluation holistique : accuracy, calibration, robustness, fairness, bias, toxicity, efficiency. Cadre méthodologique pour la définition de métriques au-delà de l'accuracy.

TruthfulQA (Lin et al., 2021) — détection des idées reçues et confabulations. Pertinent pour les misconceptions LCB-FT (seuils, exemptions, périmètre d'obligations).

2.2 Benchmarks RAG

RGB (Chen et al., 2023) et CRAG (Yang et al., Meta 2024) évaluent les systèmes RAG sur quatre capacités : robustesse au bruit (passages non pertinents dans le contexte), rejet négatif (capacité à refuser quand la réponse n'est pas dans le contexte), intégration d'information (synthèse multi-documents), robustesse contrefactuelle (résistance à des informations erronées dans le contexte). Ces axes sont directement pertinents pour évaluer un système de conformité où une hallucination peut entraîner un risque juridique réel. Ragas (Es et al., 2023) propose un ensemble de métriques opérationnelles : faithfulness, answer relevancy, context precision, context recall.

2.3 Lacune identifiée

À la connaissance des auteurs, aucun benchmark public n'évalue les systèmes d'IA sur les tâches de conformité LCB-FT en droit français ou européen. Les seuls travaux proches portent sur la détection d'opérations suspectes par apprentissage supervisé sur données transactionnelles, ce qui relève d'une problématique distincte (classification structurée, non raisonnement réglementaire). AML-Bench-FR vise à combler cette lacune en construisant un référentiel rigoureux, multi-tâches, évalué et maintenu.

3. Taxonomie des capacités évaluées

Le benchmark structure l'évaluation autour de sept capacités cognitives distinctes, chacune correspondant à un type de raisonnement mobilisé par un professionnel LCB-FT. Une question peut être taggée sur plusieurs capacités.

Code	Capacité	Description
A	Knowledge recall	Restitution exacte du corpus réglementaire : Code monétaire et financier, AMLD/AMLR, lignes directrices ACPR, doctrine TRACFIN, recommandations GAFI/FATF.
B	Statutory interpretation	Interprétation de textes ambigus, articulation de normes (lex specialis, hiérarchie UE/national, articulation avec RGPD et secret professionnel).
C	Case application (IRAC)	Qualification d'un cas concret : identification des obligations applicables, application de la règle, conclusion. Cœur du raisonnement opérationnel.
D	Risk classification	Classification du niveau de risque BC-FT (faible, standard, élevé) selon l'approche par les risques, avec justification sur la base des facteurs réglementaires.
E	Typology recognition	Détection de schémas typologiques de blanchiment ou financement du terrorisme : smurfing, layering, trade-based ML, abus d'OBNL, circuits crypto.
F	Procedural reasoning	Maîtrise des délais, formalismes, chaînes déclaratives, durées de conservation, périmètres d'information.
G	Calibration & refusal	Capacité à exprimer une incertitude calibrée, à refuser les questions hors-scope, à signaler les zones grises doctrinales et à orienter vers une expertise humaine.

3.1 Stratification croisée

Chaque entrée du dataset est en outre stratifiée selon quatre dimensions transverses :

Type d'entité assujettie : établissement de crédit, établissement de paiement, agent immobilier, notaire, expert-comptable, marchand de biens précieux (or, pierres, art), PSAN/CASP. Les obligations varient substantiellement entre ces catégories.

Juridiction : droit français (Code monétaire et financier, livre V titre VI), droit européen (AMLD 4/5/6, AMLR), normes internationales (recommandations GAFI, FATF). Permet de mesurer la confusion inter-régimes.

Temporalité : régime actuel vs régime AMLR (entrée en vigueur juillet 2027). Mesure la capacité d'anticipation du modèle.

Difficulté : calibrée empiriquement (cf. § 4.6), non auto-déclarée. Trois niveaux : easy (recall direct), medium (raisonnement à un ou deux sauts), hard (zones grises, jurisprudence Commission des sanctions, articulation de normes).

4. Méthodologie de construction du dataset

La rigueur méthodologique de construction conditionne la valeur du benchmark. Six principes structurent le processus.

4.1 Sourcing primaire

Chaque question dérive exclusivement de sources faisant autorité : textes législatifs et réglementaires (Code monétaire et financier, règlements européens), lignes directrices et recommandations des autorités (ACPR, TRACFIN, EBA, GAFI), décisions de la Commission des sanctions ACPR, jurisprudence CJUE, rapports annuels TRACFIN.

4.2 Évidence traçable

Chaque entrée du dataset stocke un champ `evidence_spans` contenant les passages exacts des sources qui justifient la réponse, avec référence officielle (article, considérant, paragraphe, URL). Cette traçabilité poursuit deux objectifs : permettre l'évaluation séparée du retrieval et assurer l'auditabilité juridique des réponses de référence.

4.3 Validation contradictoire experte

Le processus d'annotation suit un protocole en quatre étapes inspiré de GPQA :

1. Rédaction initiale par un auteur expert LCB-FT.
2. Réponse en aveugle par un second expert. Tout désaccord déclenche une discussion documentée et une révision.
3. Validation finale par un troisième expert sur la version révisée.
4. Mesure d'inter-annotator agreement (kappa de Cohen, kappa de Fleiss) sur un sous-ensemble de 100 questions annotées indépendamment par trois experts. Une catégorie présentant un kappa inférieur à 0,6 est jugée mal définie et reformulée.

4.4 Design adversarial

Pour une fraction substantielle des questions (cible : 25 %), une variante adversariale est produite. La variante modifie un seul paramètre du cas (juridiction, type d'entité assujettie, seuil voisin, période d'application) de manière à inverser ou modifier la réponse, tout en conservant une formulation très proche.

4.5 Anti-contamination

Quatre parades sont mises en œuvre : (i) **held-out privé** — 30 % des questions ne sont jamais publiées et constituent l'ensemble d'évaluation officielle accessible uniquement via une API ; (ii) **canary strings**

— chaque entrée publique contient un identifiant unique permettant la détection a posteriori d'une contamination ; (iii) **versionnage et rotation** — publication par versions semestrielles avec rotation partielle ; (iv) **reformulation** — les cas réels sont systématiquement reformulés et anonymisés.

4.6 Calibration empirique de la difficulté

La difficulté n'est pas auto-déclarée. Elle est calibrée empiriquement à partir des taux de réussite d'un panel de modèles de référence (cible initiale : Claude Opus 4.7, GPT-5, Gemini 2.5 Pro, Mistral Large, Llama 70B). Les questions sont réparties en trois quantiles. Le recours à un modèle d'Item Response Theory (IRT) est envisagé en v2.

5. Formats de tâches

AML-Bench-FR adopte une diversité de formats : QCM (4-5 options), réponses ouvertes courtes (exact match / regex), réponses ouvertes longues (rubrique structurée + juge LLM), classification multi-label (F1-score), ranking (Kendall's tau), génération structurée (fiche KYC, trame DS), tâches de refus (taux de fausse confiance), tâches contrefactuelles.

6. Métriques d'évaluation

Conformément à HELM, l'évaluation est holistique : **accuracy** ventilée par capacité / entité / difficulté ; **calibration** (Expected Calibration Error) ; **faithfulness /** hallucination rate ; **citation accuracy** ; **refusal appropriateness** (matrice de confusion sur questions à refus attendu) ; **robustness** sous perturbations ; **self-consistency** sur n=5 runs ; **métriques RAG** (Context Precision, Context Recall) ; **coût et latence**.

7. Protocole d'évaluation

Few-shot uniforme (0-shot, 3-shot, 5-shot avec exemples identiques) ; templates de prompt figés et publiés ; seeds fixes et températures reportées (T=0 pour le score officiel, T=0,7 pour la mesure de variance) ; multiplicité des runs (n=3 minimum) ; baseline humaine sur sous-ensemble représentatif ; baselines triviales (random, majority) pour les classifications.

8. Gouvernance, diffusion et soutenabilité

Documentation. Datasheet for Datasets (Gebru et al., 2018), Model Cards (Mitchell et al., 2019).

Licence et accès. Partie publique sous CC-BY 4.0. Held-out sous EULA stricte, accessible uniquement via API d'évaluation.

Leaderboard et soumission. Hugging Face Spaces, soumission via API, exécution sur held-out contrôlé.

Publication scientifique. Article rédigé en parallèle, dépôt arXiv puis soumission à un workshop NLP-Law (NLLP, Jurix) ou aux findings d'une conférence de référence (ACL, EMNLP). Co-auteurship académique recherchée.

Maintenance. Versionnage sémantique. Rotation partielle à chaque version mineure ; refonte taxonomique à chaque version majeure. Cadence semestrielle.

9. Roadmap

Phase	Durée	Livrables
0 — Cadrage *(actuel)*	2-3 semaines	Revue de littérature consolidée. Position paper finalisé. Premiers contacts académiques. Définition technique du schéma de données.
1 — Pilote	1-2 mois	50 questions sur une vertical (proposition : agent immobilier). Évaluation de 4-5 modèles. Mini-rapport public.
2 — Extension	3-4 mois	500-800 questions couvrant la taxonomie complète. Recrutement de 2-3 co-annotateurs experts. Premier leaderboard public. Soumission arXiv.
3 — Diffusion	Continue	Maintenance du leaderboard. Versions semestrielles. Soumission conférence. Partenariats institutionnels (ACPR, AMLA). Extension multi-juridictionnelle (BE, LU, DE).

10. Risques et limites

Risque méthodologique. Une taxonomie imparfaite ou un sourcing partiel biaiseraient les conclusions. Atténuation : validation contradictoire experte, mesure formelle d'inter-annotator agreement, ouverture du processus à la revue par les pairs.

Risque de contamination et d'obsolescence. La diffusion publique érode la valeur diagnostique. Atténuation : held-out privé, canary strings, rotation versionnée. L'évolution réglementaire (AMLR 2027) impose en outre une maintenance active.

Risque d'instrumentalisation. Un benchmark peut être utilisé pour justifier abusivement le déploiement d'un système IA en production critique. Atténuation : la documentation insiste explicitement sur le fait que le benchmark mesure des capacités circonscrites et ne constitue ni une certification, ni une garantie de conformité opérationnelle.

11. Conclusion et appel à collaboration

AML-Bench-FR se positionne à l'intersection d'un besoin réglementaire émergent, d'une lacune scientifique avérée et d'une opportunité industrielle réelle. Le projet recherche dès la phase 0 des collaborations : (i) un partenariat académique avec un laboratoire de NLP juridique pour la co-auteurship scientifique ; (ii) un panel d'experts LCB-FT praticiens pour l'annotation contradictoire ; (iii) un dialogue avec les autorités (ACPR, AMLA) pour aligner le benchmark sur les futures attentes supervisoires.

Contact : research@lutteblanchiment.fr — formulaire en ligne sur https://lutteblanchiment.fr/research/aml-bench-fr#collab.

Références

Chen, J. et al. (2023). *Benchmarking Large Language Models in Retrieval-Augmented Generation*. arXiv:2309.01431.

Es, S. et al. (2023). *RAGAS: Automated Evaluation of Retrieval Augmented Generation*. arXiv:2309.15217.

Gebru, T. et al. (2018). *Datasheets for Datasets*. arXiv:1803.09010.

Guha, N. et al. (2023). *LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models*. arXiv:2308.11462.

Hendrycks, D. et al. (2020). *Measuring Massive Multitask Language Understanding*. arXiv:2009.03300.

Islam, P. et al. (2023). *FinanceBench: A New Benchmark for Financial Question Answering*. arXiv:2311.11944.

Liang, P. et al. (2022). *Holistic Evaluation of Language Models (HELM)*. arXiv:2211.09110.

Lin, S. et al. (2021). *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. arXiv:2109.07958.

Mitchell, M. et al. (2019). *Model Cards for Model Reporting*. arXiv:1810.03993.

Rein, D. et al. (2023). *GPQA: A Graduate-Level Google-Proof Q&A Benchmark*. arXiv:2311.12022.

Yang, X. et al. (2024). *CRAG — Comprehensive RAG Benchmark*. Meta AI.

Sources institutionnelles : Code monétaire et financier (livre V, titre VI) ; Directives AMLD 4/5/6 ; Règlement AMLR ; Lignes directrices ACPR et TRACFIN ; Recommandations GAFI/FATF ; Décisions de la Commission des sanctions ACPR ; Rapports annuels TRACFIN.