Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Évaluation des performances des systèmes à base de LLM : métriques globales et locales au service de l'adoption

Jean-Baptiste Juin[1] Thomas Leguere[1]
(1) Cross Data, 25 Rue Lenepveu, 49100 Angers
jbjuin@crossdata.tech, tleguere@crossdata.tech

RÉSUMÉ

L'évaluation des systèmes à base de grands modèles de langage (LLM) en contexte applicatif reste un défi ouvert : les sorties génératives sont difficiles à évaluer objectivement, et les systèmes de recherche sémantique dense ne disposent d'aucun mécanisme d'abstention natif. Nous présentons *llm-app-metrics*, un cadre méthodologique unifié fondé sur la comparaison de distributions de scores entre prédictions correctes (ρ^+) et incorrectes (ρ^-). Ce principe s'applique à la confiance en classification (logprobs) comme à la confiance en retrieval (scores cosinus). Un pipeline bayésien complet permet de produire un modèle de confiance calibré et sérialisable, intégrable en production. Les expérimentations sur des benchmarks publics (TREC-DL, mMARCO) et des données de production réelles valident la séparabilité des distributions et l'opérationnalité du pipeline.

ABSTRACT

Evaluating LLM-based System Performance : Global and Local Metrics for Trustworthy Adoption

Evaluating Large Language Model (LLM)-based systems in real-world applications remains an open challenge : generative outputs are difficult to assess objectively, and dense semantic retrieval systems lack a native abstention mechanism. We introduce *llm-app-metrics*, a unified methodological framework based on comparing score distributions between correct (ρ^+) and incorrect (ρ^-) predictions. This principle applies to both classification confidence (logprobs) and retrieval confidence (cosine scores). A complete Bayesian pipeline produces a calibrated and serializable confidence model ready for production deployment. Experiments on public benchmarks (TREC-DL, mMARCO) and real production data validate the separability of distributions and the operational viability of the pipeline.

MOTS-CLÉS : LLM, RAG, métriques de confiance, distributions de scores, données synthétiques, recherche d'information.

KEYWORDS: LLM, RAG, confidence metrics, score distributions, synthetic data, information retrieval.

1 Introduction

L'intégration de modèles de langage de grande taille (LLM) au sein d'applications métier – systèmes de génération augmentée par la recherche (RAG), assistants conversationnels, pipelines de traitement automatique du langage naturel – soulève une question fondamentale : comment mesurer objectivement la qualité des réponses produites par ces systèmes ? Contrairement aux modèles d'apprentissage supervisé classiques, évaluables par des métriques bien établies (précision, rappel, F1-score) sur des

jeux annotés, les sorties des LLMs sont génératives, contextuelles et difficiles à circonscrire (Kadavath *et al.*, 2022; Desai & Durrett, 2020). L'absence de référentiel de vérité terrain universel, la variabilité des formulations acceptables pour une même réponse correcte, et la sensibilité aux choix d'ingénierie du prompt rendent l'évaluation particulièrement complexe.

Cette difficulté est amplifiée par un problème structurel de la recherche sémantique par vecteurs denses, au cœur de la plupart des systèmes RAG. Ces approches représentent documents et requêtes sous forme de vecteurs dans un espace sémantique de haute dimension et utilisent une distance cosinus pour classer les documents par pertinence. Or, cette mécanique produit *toujours* un top-k de résultats, même si aucun document n'est pertinent pour la requête donnée. L'absence d'un mécanisme d'abstention – la capacité de signaler qu'aucun résultat fiable n'a été trouvé – constitue un problème fondamental pour la confiance utilisateur (Rossi *et al.*, 2024; Arampatzis *et al.*, 2009). Sans seuil de pertinence fiable, le système RAG peut injecter dans le contexte du LLM des documents non pertinents, générant des réponses faussement confiantes et factuellement incorrectes.

Cette problématique se décline sur deux niveaux complémentaires. D'une part, des métriques *globales* sont nécessaires pour guider les équipes de développement : comparer des architectures, évaluer l'impact d'un changement de modèle, valider qu'une nouvelle version progresse sur les cas cibles. D'autre part, des métriques *locales* – attachées à chaque réponse individuelle – sont indispensables pour que l'utilisateur final puisse exercer son jugement critique : un indice de confiance, une indication des sources, ou un signal d'incertitude constituent autant de leviers pour éviter une surconfiance aveugle dans le système.

Ce constat est issu d'expériences terrain menées par Cross Data, société spécialisée en conseil et intégration d'intelligence artificielle, sur plusieurs projets clients mobilisant des LLMs dans des contextes exigeants : extraction d'entités nommées depuis des documents hétérogènes, analyse de conformité réglementaire, assistant conversationnel RAG sur corpus spécialisé. La convergence de ces besoins a conduit à formaliser, à partir d'avril 2025, un projet de R&D transversal baptisé *llm-app-metrics*.

Question de recherche. *Comment construire un cadre méthodologique unifié et opérationnel permettant de fournir, pour tout système applicatif à base de LLM, des métriques de confiance locales calibrées sur le comportement réel du système dans son domaine applicatif ?*

Contributions. Ce travail apporte deux contributions principales : (1) un principe méthodologique unificateur – la comparaison de distributions ρ^+ (scores des prédictions correctes) et ρ^- (scores des prédictions incorrectes) – matérialisé dans un pipeline de confiance bayésien opérationnel, de la calibration à la sérialisation d'un modèle intégrable en production ; (2) une méthodologie de génération et de validation de données synthétiques, incluant une annotation humaine partielle et un filtrage automatique, permettant de calibrer ce modèle de confiance en l'absence de jeu de données annoté exhaustif.

L'article est organisé comme suit : la section 2 présente l'état de l'art, la section 3 détaille la méthodologie, la section 4 rapporte les expérimentations et résultats, et la section 5 conclut avec les perspectives de recherche.

2 État de l'art

Le projet *llm-app-metrics* se situe à la convergence de plusieurs axes de recherche. Nous présentons les travaux sur lesquels notre méthodologie s'appuie et identifions les limites qui motivent notre contribution.

2.1 Modélisation des distributions de scores en recherche d'information

La modélisation statistique des distributions de scores en IR est un domaine actif depuis les travaux fondateurs de Swets (Swets, 1963). Pour les distributions issues de méthodes de recherche éparse, le modèle dominant repose sur une exponentielle négative pour les documents non pertinents et une gaussienne pour les documents pertinents. Manmatha et al. (Manmatha *et al.*, 2001) ont montré que ce modèle permet de convertir des scores de recherche en probabilités de pertinence via le théorème de Bayes, validant cette approche sur les données TREC avec des moteurs variés (INQUERY, SMART, LSI). Kanoulas et al. (Kanoulas *et al.*, 2010) ont approfondi cette analyse en dérivant mathématiquement les distributions de scores attendues pour BM25 et les modèles de langue, justifiant l'utilisation de mélanges de gaussiennes. Arampatzis et al. (Arampatzis *et al.*, 2009) ont développé des modèles de distributions tronquées pour l'optimisation de seuils, avec des résultats significativement meilleurs sur TREC Legal Track.

Ces travaux portent cependant exclusivement sur des moteurs de recherche classiques (BM25, TF-IDF). Les systèmes RAG modernes reposent sur des encodeurs denses dont les distributions de scores cosinus présentent des caractéristiques différentes – notamment des distributions plus concentrées et des queues moins marquées. Notre contribution consiste à transposer cette approche au cas des scores de similarité sémantique dense, en testant systématiquement une dizaine de lois candidates plutôt que de présupposer un modèle spécifique.

2.2 Abstention en recherche sémantique dense

Un problème fondamental des systèmes de recherche par vecteurs denses est l'absence de mécanisme d'abstention naturel : contrairement à la recherche lexicale, la recherche dense retourne toujours un top-k de résultats, même lorsqu'aucun document n'est pertinent. Rossi et al. (Rossi *et al.*, 2024) ont proposé un « Cosine Adapter » – une fonction de transformation apprise, dépendante de la requête, qui convertit les scores cosinus en scores de pertinence interprétables. Leurs expériences sur MS MARCO et des données internes montrent une amélioration significative de la précision. Notre approche partage cet objectif mais diffère par la méthode : plutôt qu'un adaptateur neuronal nécessitant un ré-entraînement par domaine, nous proposons une sélection automatique de modèle parmi une famille de distributions candidates, évaluées par le critère de log-vraisemblance ($\log \mathcal{L}$).

2.3 Estimation de l'incertitude des LLMs

Kadavath et al. (Kadavath *et al.*, 2022) ont montré que les probabilités de sortie des LLMs sont partiellement corrélées à la justesse de leurs réponses, mais Desai & Durrett (Desai & Durrett, 2020) ont démontré que ces probabilités sont mal calibrées : elles ne reflètent pas fidèlement la probabilité

réelle de correction, nécessitant des méthodes de calibration post-hoc. Ma et al. (Ma *et al.*, 2025) ont confirmé que la normalisation softmax fait perdre l'information de force d'évidence accumulée durant l'entraînement. Notre approche est complémentaire : plutôt que d'estimer l'incertitude intrinsèque du modèle, nous calibrons empiriquement les logprobs en construisant les distributions ρ^+ et ρ^- à partir de données annotées – une calibration extrinsèque plus robuste car elle ne fait aucune hypothèse sur la signification absolue des probabilités du modèle.

2.4 Génération de données synthétiques pour l'évaluation

L'absence de données annotées dans les domaines spécifiques est un frein majeur à l'évaluation des systèmes de recherche d'information. Bonifacio et al. (Bonifacio *et al.*, 2022) ont proposé InPars, utilisant les capacités *few-shot* des LLMs pour générer des requêtes synthétiques à partir de documents. Husain (Husain, 2024) a popularisé l'approche par dimensions pour contrôler systématiquement les axes de variation des questions générées (type de tâche, complexité, style). Notre méthodologie s'appuie sur ces travaux et intègre la génération synthétique directement dans le pipeline de calibration de confiance – un usage original par rapport à l'entraînement de modèles de retrieval.

2.5 Positionnement

Le tableau 1 résume le positionnement de notre travail. Notre contribution se distingue par l'unification de ces axes dans un cadre méthodologique cohérent : la comparaison de distributions ρ^+ vs ρ^- est le principe transversal qui s'applique tant à la confiance en classification (logprobs) qu'à la confiance en retrieval (scores cosinus). Cette unification constitue, à notre connaissance, une approche qui n'a pas d'équivalent direct dans la littérature.

Axe	**Approche existante**	**Notre contribution**
Distributions de scores IR	Modèles exp.-gaussiens sur BM25/TF-IDF (Manmatha *et al.*, 2001; Kanoulas *et al.*, 2010)	Extension aux scores denses, sélection automatique parmi 10 familles
Abstention en retrieval dense	Cosine Adapter appris (Rossi *et al.*, 2024), seuils tronqués (Arampatzis *et al.*, 2009)	Seuil bayésien calibré par distributions, sans ré-entraînement
Incertitude LLM	LogTokU (Ma *et al.*, 2025), calibration post-hoc (Desai & Durrett, 2020)	Calibration extrinsèque par ρ^+/ρ^-, indépendante du modèle
Données synthétiques	InPars (Bonifacio *et al.*, 2022), dimensions (Husain, 2024)	Usage pour calibration de confiance

TABLE 1 – Positionnement par rapport à l'état de l'art.

3 Méthodologie

La méthodologie *llm-app-metrics* repose sur un principe transversal : la comparaison de distributions de scores entre prédictions correctes (ρ^+) et incorrectes (ρ^-). Ce principe s'applique de manière unifiée à la confiance en classification par LLM (logprobs) et à la confiance en retrieval (scores de similarité cosinus). Nous détaillons ci-dessous les deux composants principaux de la méthodologie : le pipeline de confiance bayésien et la génération de données synthétiques.

3.1 Pipeline de confiance bayésien

3.1.1 Principe

Soit un système produisant, pour chaque prédiction, un score $s \in \mathbb{R}$ reflétant un degré de certitude : probabilité logprob d'un token de classification, ou score de similarité cosinus en retrieval. À partir d'un jeu de données annoté, deux distributions sont construites : ρ^+, la distribution de s dans les cas où le système a prédit correctement, et ρ^-, la distribution de s dans les cas où le système s'est trompé.

Pour une nouvelle prédiction de score s, la probabilité postérieure bayésienne de correction est :

$$P(\text{positif} \mid s) = \frac{f^+(s) \cdot \pi^+}{f^+(s) \cdot \pi^+ + f^-(s) \cdot \pi^-} \tag{1}$$

où f^+ et f^- sont les densités de probabilité ajustées sur ρ^+ et ρ^- respectivement, et π^+, π^- les priors (proportions de prédictions correctes et incorrectes dans le contexte d'utilisation). Ce score de confiance, compris dans $[0, 1]$, est directement interprétable par l'utilisateur final. En pratique, le prior est estimé à partir du ratio observé sur les données de calibration ; la sensibilité de ce choix est discutée en section 4.3.

3.1.2 Ajustement et sélection automatique de distributions

Le choix du modèle de distribution est crucial : un mauvais ajustement conduit à des scores de confiance mal calibrés. Plutôt que de présupposer un modèle paramétrique spécifique – comme les travaux classiques en IR qui utilisent un couple exponentielle-gaussienne (Manmatha *et al.*, 2001) – nous testons systématiquement dix familles de distributions candidates sur les scores observés : deux estimateurs à noyaux (KDE avec bandes passantes de Scott et Silverman), quatre mélanges gaussiens (GMM à $K = 2, 3, 4, 5$ composantes) et quatre distributions paramétriques (Normale, Log-Normale, Beta, Gamma).

La sélection du meilleur modèle est réalisée automatiquement en comparant les candidats par la métrique de log-vraisemblance $\log \mathcal{L}$. Cette approche permet de s'adapter aux caractéristiques spécifiques de chaque combinaison corpus/encodeur sans intervention manuelle (figure 1).

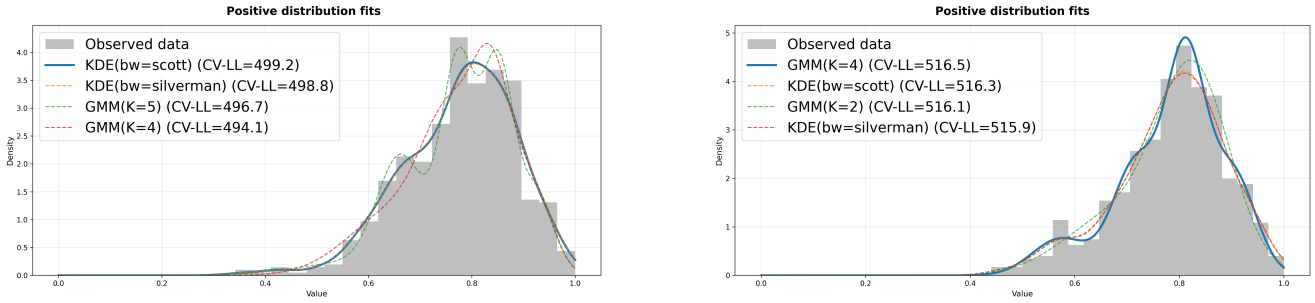

FIGURE 1 – Distributions ρ^+ sur un jeu de données interne de 470 entrées. Gauche : *dangvantuan/sentence-camembert-large*. Droite : *Qwen/Qwen3-Embedding-0.6B*.

3.1.3 Pipeline complet

Le pipeline opérationnel procède en cinq étapes : (1) normalisation des scores de distance sémantique en scores $\in [0, 1]$; (2) ajustement des 10 modèles candidats sur les scores positifs et négatifs ; (3) sélection automatique du meilleur ajustement par critère $\log \mathcal{L}$; (4) calcul de la probabilité postérieure bayésienne $P(\text{positif} \mid s)$ pour chaque score observé ; (5) sérialisation du modèle de confiance complet (distributions ajustées, prior, métadonnées) en fichier JSON réutilisable en production. Le modèle sérialisé peut être rechargé sans recalcul, permettant à tout projet disposant d'un corpus documentaire de produire son propre modèle de confiance calibré.

3.2 Génération de données synthétiques

La construction des distributions ρ^+ et ρ^- nécessite un jeu de données annoté (paires question/document avec jugements de pertinence). Or, dans les contextes applicatifs réels, de tels jeux de données sont quasi systématiquement absents (Bonifacio *et al.*, 2022). La génération de données synthétiques est donc un prérequis fondamental à l'application de la méthodologie.

3.2.1 Génération de paires question-document

Pour chaque chunk du corpus documentaire, un LLM génère une question dont la réponse peut être trouvée dans ce chunk, créant par construction une paire positive (q_i, d_i), tandis que toutes les autres combinaisons $(q_i, d_j)_{j \neq i}$ constituent des paires négatives. Le fichier de jugements de pertinence (*qrels*) correspondant est produit automatiquement au format standard *ir_measures*.

3.2.2 Validation des questions synthétiques

La validation des questions synthétiques est un enjeu important pour assurer la qualité de la mesure des performances. Nous proposons une validation en deux étapes : (1) construction d'un ensemble de questions annotées manuellement, (2) évaluation de la qualité des questions synthétiques par deux méthodes complémentaires – un jury de LLMs et une comparaison géométrique des embeddings.

Un jeu de référence annoté (*golden dataset*) est d'abord constitué sur un sous-ensemble des chunks. Deux informations sont récupérées auprès des annotateurs : (a) une évaluation binaire de la question

synthétique associée au chunk, (b) une question alternative proposée par l'annotateur.

La première méthode de validation (A) repose sur un jury de LLMs, utilisé pour l'évaluation binaire des questions synthétiques et annotées. Les votes du jury, combinés aux votes des annotateurs, permettent à la fois de valider la fiabilité du jury et d'évaluer la qualité des questions synthétiques. Une fois validé, le jury est utilisé pour évaluer l'ensemble des questions synthétiques ; les questions rejetées par vote majoritaire simple sont écartées du jeu de données d'évaluation. Une optimisation de la partie *few-shot learning* du prompt du jury est réalisée, en maximisant la conformité aux évaluations des annotateurs (Khattab *et al.*, 2024, 2022).

Etape 2 - méthode B : les échantillons de questions synthétiques et de questions annotées sont comparées en utilisant les propriétés géométriques de leurs embeddings. Dans un premier temps nous comparons la distribution des distances (cosinus) entre les questions synthétiques et les questions annotées correspondantes (même chunk - paires positives) avec la distribution des distances entre questions synthétiques sur l'ensemble des chunks (paires négatives). Dans un seconde temps nous compléterons cette analyse avec les métriques décrites dans (Naeem *et al.*, 2020) et (Kynkäänniemi *et al.*, 2019). Les métrique de précision et rappel (Kynkäänniemi) ainsi que la densité et la couverture (Naeem) seront utilisées pour évaluer la qualité des questions synthétiques vis-à-vis des questions annotées. La métrique la plus pertinente étant la "Coverage" définie dans Naeem : on veut être sûrs que les questions annotées soient bien représentées (couvertes) par les questions synthétiques.

4 Expérimentations et résultats

Les expérimentations s'appuient sur des benchmarks publics de référence en recherche d'information et sur des données de production issues de projets clients (tableau 2).

Expérimentation	**Type**	**Source**
Métriques de retrieval	Benchmark	MS MARCO (Nguyen *et al.*, 2016), mMARCO/fr (Bonifacio *et al.*, 2021), TREC-DL 2019/2021
Pipeline de confiance	Synth. + public	GPT-4o-mini + mMARCO/fr (1 000 docs)
Classification	Production	Projet client (documents juridiques)
Génération synth.	Synthétique	Corpus interne (documents de test)

TABLE 2 – Synthèse des données utilisées.

4.1 Séparabilité des distributions sur scores denses

L'hypothèse fondatrice est que les distributions de scores de similarité cosinus pour les paires pertinentes (ρ^+) et non pertinentes (ρ^-) sont effectivement séparables lorsque le modèle d'embedding est adapté au domaine. Sur un sous-ensemble de 1 000 documents de mMARCO/fr, le pipeline a été exécuté avec deux modèles d'embedding : *openai/text-embedding-3-small* et *dangvantuan/sentence-camembert-large*. Avec le modèle OpenAI, la séparation est nette (figure 2, gauche) et le seuil bayésien optimal est $\theta = 0{,}830$. Avec CamemBERT, les paires positives ($\sim 0{,}85$) chevauchent davantage les

négatives ($\sim 0{,}5$), rendant la calibration plus délicate (figure 2, droite). Cette dépendance au modèle d'embedding est discutée en section 4.3.

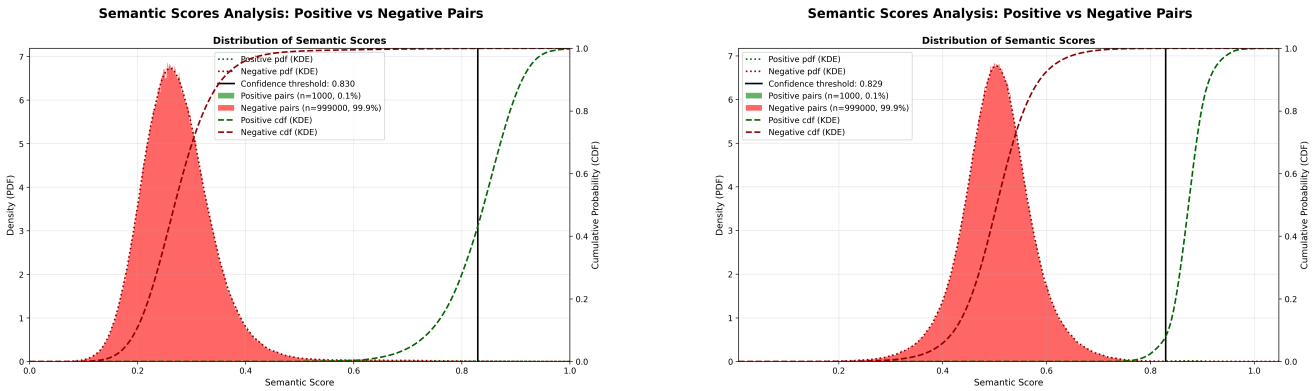

FIGURE 2 – Distributions ρ^+ et ρ^- sur mMARCO/fr. Gauche : *openai/text-embedding-3-small* ($\theta = 0{,}830$). Droite : *dangvantuan/sentence-camembert-large*, chevauchement plus marqué.

4.1.1 Comportement du pipeline sur des requêtes humaines

Afin d'évaluer le comportement du modèle de confiance face à des requêtes réalistes, nous avons soumis un ensemble de questions formulées manuellement au pipeline calibré sur mMARCO/fr. Ces questions se répartissent en deux catégories : des requêtes *en domaine*, dont la réponse est attendue dans le corpus, et des requêtes *hors domaine*, volontairement sans rapport avec le contenu indexé. L'objectif est de vérifier que le score de confiance bayésien $P(\text{positif} \mid s)$ discrimine correctement ces deux cas – en particulier que les requêtes hors domaine se voient attribuer un score de confiance faible, malgré le fait que le système de retrieval retourne nécessairement un résultat.

Les requêtes en domaine obtiennent des scores de confiance élevés, confirmant que le pipeline identifie correctement la présence de documents pertinents. À l'inverse, les requêtes hors domaine reçoivent des scores de confiance faibles malgré des scores de similarité cosinus non nuls – illustrant précisément le mécanisme d'abstention que le pipeline vise à fournir. Cet écart entre score brut et score calibré valide l'apport de la modélisation bayésienne par rapport à un seuillage direct sur la similarité cosinus.

4.2 Validation de la génération synthétique

La méthodologie de génération a été appliquée sur un corpus documentaire interne. Un jeu de données synthétiques structuré et diversifié a été produit, permettant d'évaluer les performances du système RAG en l'absence de tout jeu de données annoté préexistant.

4.2.1 Filtrage par jury de LLMs

Afin de valider la qualité des questions générées, une étape de filtrage par jury de LLMs a été mise en place. Quatre modèles – Mistral-small-4-thinking, Qwen-3.6-27B, Gemma-4-31-A4B et Gemma-4-26B – évaluent indépendamment si la question synthétique est pertinente par rapport au

chunk considéré. Seules les paires obtenant un vote unanime des quatre juges sont conservées, ce qui a permis de constituer un jeu de données filtré de 470 paires positives et 589 paires négatives. Sur les données métier le Jury atteint 92.7 % d'accuracy et sur mMARCO/fr le Jury atteint 95.46 % d'accuracy sur l'identification de l'adéquation des questions sur des paires positives (question générée à partir du chunk) et négatives (question et chunk choisis aléatoirement). Cette méthode permet de fournir un premier indicateur de la qualité des questions générées.

4.2.2 Comparaison géométrique avec les questions humaines

Afin de mesurer l'écart entre les questions synthétiques et des requêtes réelles, un ensemble de questions formulées par des utilisateurs a été collecté sur le même corpus. Les embeddings des questions synthétiques et des questions humaines ont été projetés dans un espace réduit (figure 3). Chaque question synthétique est associée à la question humaine portant sur le même document, matérialisant ainsi les écarts de formulation à intention constante.

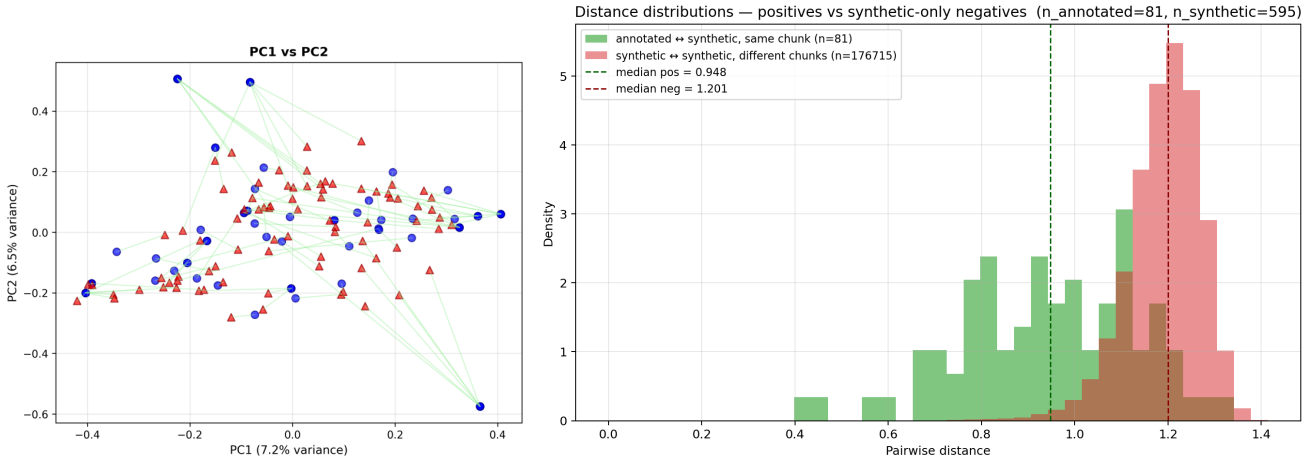

FIGURE 3 – (Gauche) Projection des embeddings des questions synthétiques (bleu) et des questions humaines (rouge) sur le même corpus. Les traits verts relient chaque paire (synthétique, humaine) portant sur le même document. La distance entre les points d'une même paire reflète l'écart de formulation entre la génération automatique et l'usage réel. (Droite) Comparaison des distributions des distances cosinus entre paires positives (annotée / synthétique - même chunk) et des distances entre paires négatives (synthétique / synthétique) sur l'ensemble des autres chunks.

Les résultats montrent que les questions synthétiques et humaines occupent des régions partiellement chevauchantes de l'espace sémantique, ce qui confirme la pertinence globale de la génération. Cependant, les distances intra-paires (traits verts) révèlent une variabilité notable : certaines questions synthétiques sont proches de leur équivalent humain, tandis que d'autres s'en éloignent significativement, traduisant des différences de formulation, de niveau de détail ou d'ambiguïté. Cette observation motive les travaux en cours sur l'alignement des distributions synthétiques et réelles par optimisation de prompts (cf. section 5). La détection automatique de paires divergentes peut permettre une analyse métier pertinente en soit pour la compréhension des usages et l'optimisation des prompts.

4.3 Limites identifiées

Nos expérimentations ont mis en évidence plusieurs limites qu'il convient de discuter.

La sensibilité au modèle d'embedding est la plus notable : les distributions de scores et les seuils optimaux varient significativement d'un encodeur à l'autre, comme l'illustre la comparaison entre *openai/text-embedding-3-small* et *dangvantuan/sentence-camembert-large*. Un modèle de confiance calibré sur un encodeur n'est pas transférable à un autre, ce qui impose une recalibration systématique lors de tout changement de modèle. En production la surveillance statistique des scores et la comparaison systématique aux courbres de calibration peut permettre le déclenchement d'une recalibration automatique pour pallier à d'éventuelles dérives du corpus et des usages.

La représentativité des données synthétiques constitue une hypothèse forte non encore validée quantitativement sur des corpus diversifiés. Les questions générées par un LLM tendent à être plus explicites et mieux formulées que les requêtes réelles des utilisateurs, qui peuvent être ambiguës, contenir des fautes ou combiner plusieurs intentions. Le décalage entre les distributions de scores synthétiques et réelles pourrait affecter la qualité de la calibration. En production il sera essentiel de collecter les questions des utilisateurs et de systématiquement réévaluer la qualité des questions synthétiques en mesurant la dérives des distributions de distances entre paires positives vis à vis des distributions de référence. Par ailleurs les résultats obtenus par Jury de LLMs sur l'évaluation de la pertinence des questions synthétiques devront être confirmés en validant les analyses faites sur des données annotées.

Enfin, le calcul de $P(\text{positif} \mid s)$ via l'équation (1) nécessite la connaissance du prior π^+. Sur les données de calibration synthétiques, ce ratio est connu par construction. En production, il est inconnu et variable : il dépend de la richesse du corpus par rapport à la requête et du domaine applicatif. Nous utilisons actuellement le prior estimé sur les données de calibration, mais la recherche d'une méthode d'estimation plus robuste en conditions réelles reste un problème ouvert (Manmatha *et al.*, 2001). L'usage de techniques comme le "Platt scaling" (Platt, 2000) sera à étudier notamment leur validité en dehors des cas de classification.

5 Conclusion et perspectives

Nous avons présenté *llm-app-metrics*, un cadre méthodologique unifié pour l'évaluation des systèmes à base de LLM, fondé sur le principe de comparaison de distributions ρ^+/ρ^-. Ce cadre s'applique à la confiance en classification (logprobs) comme à la confiance en retrieval (scores cosinus). Les expérimentations sur des benchmarks publics (TREC-DL, mMARCO) et sur des données de production réelles ont validé l'hypothèse fondatrice : les distributions de scores des encodeurs denses sont effectivement séparables, permettant le calcul d'un seuil bayésien de confiance interprétable.

Plusieurs limites ont été identifiées – sensibilité au modèle d'embedding, représentativité des données synthétiques, estimation du prior en production – qui orientent les travaux en cours. À court terme, nous explorons l'alignement des questions synthétiques sur un échantillon de référence de requêtes réelles, via des techniques de réduction de dimension (UMAP, t-SNE) des métriques adaptées du domaine de la vision (Naeem *et al.*, 2020) et d'optimisation automatique de prompts. À plus long terme, deux pistes sont envisagées : d'une part, l'extension du cadre ρ^+/ρ^- à la pondération d'un jury de LLMs évaluant la qualité de génération, en calibrant la confiance de chaque juré via ses logprobs

de décision (Zheng *et al.*, 2023; Verga *et al.*, 2024; Tian *et al.*, 2025) ; d'autre part, la distillation de ces évaluations dans des modèles légers (fine-tuning LoRA, classifieurs sur embeddings) compatibles avec un usage en production temps réel.

Références

ARAMPATZIS A., KAMPS J. & ROBERTSON S. (2009). Where to stop reading a ranked list ? threshold optimization using truncated score distributions. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 524–531. DOI : 10.1145/1571941.1572031.

BONIFACIO L., ABONIZIO H., FADAEE M. & NOGUEIRA R. (2021). mMARCO : A multilingual version of the MS MARCO passage ranking dataset. arXiv : 2108.13897.

BONIFACIO L., ABONIZIO H., FADAEE M. & NOGUEIRA R. (2022). InPars : Data augmentation for information retrieval using large language models. arXiv : 2202.05144.

DESAI S. & DURRETT G. (2020). Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 295–302. DOI : 10.18653/v1/2020.emnlp-main.21.

HUSAIN H. (2024). What is the best approach for generating synthetic data ? Blog post : hamel.dev/blog/posts/evals-faq.

KADAVATH S., CONERLY T., ASKELL A., HENIGHAN T., DRAIN D., PEREZ E., SCHIEFER N., HATFIELD-DODDS Z., DASSARMA N., TRAN-JOHNSON E. *et al.* (2022). Language models (mostly) know what they know. arXiv : 2207.05221.

KANOULAS E., DAI K. & ASLAM J. A. (2010). Score distribution models : Assumptions, intuition, and robustness to score manipulation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 242–249. DOI : 10.1145/1835449.1835491.

KHATTAB O., SANTHANAM K., LI X. L., HALL D., LIANG P., POTTS C. & ZAHARIA M. (2022). Demonstrate-search-predict : Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv :2212.14024*.

KHATTAB O., SINGHVI A., MAHESHWARI P., ZHANG Z., SANTHANAM K., VARDHAMANAN S., HAQ S., SHARMA A., JOSHI T. T., MOAZAM H., MILLER H., ZAHARIA M. & POTTS C. (2024). Dspy : Compiling declarative language model calls into self-improving pipelines.

KYNKÄÄNNIEMI T., KARRAS T., LAINE S., LEHTINEN J. & AILA T. (2019). Improved precision and recall metric for assessing generative models.

MA H., ZHU Y., WANG P., LIU Y. & SHEN X. (2025). Estimating LLM uncertainty with evidence. arXiv : 2502.00290.

MANMATHA R., RATH T. & FENG F. (2001). Modeling score distributions for combining the outputs of search engines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 267–275. DOI : 10.1145/383952.384005.

NAEEM M. F., OH S. J., UH Y., CHOI Y. & YOO J. (2020). Reliable fidelity and diversity metrics for generative models.

NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R. & DENG L. (2016). MS MARCO : A human generated MAchine reading COmprehension dataset. In *Proceedings of the Workshop on Cognitive Computation : Integrating Neural and Symbolic Approaches (CoCo@NIPS)*. arXiv : 1611.09268.

PLATT J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, **10**.

ROSSI N., JHA S. & PAPPU S. R. (2024). Relevance filtering for embedding-based retrieval. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*. DOI : 10.1145/3627673.3680082.

SWETS J. A. (1963). Information retrieval systems. *Science*, **141**(3577), 245–250. DOI : 10.1126/science.141.3577.245.

TIAN Z., HU R., WANG H., WANG S. & CHE W. (2025). Overconfidence in LLM-as-a-judge : Diagnosis and confidence-driven solution. arXiv : 2508.06225.

VERGA P., HOFSTÄTTER S., ALTHAMMER S., SU Y., GUPTA A., DIBIA V. & MOSCHITTI A. (2024). Replacing judges with juries : Evaluating LLM generations with a panel of diverse models. arXiv : 2404.18796.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. P. *et al.* (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv : 2306.05685.