Benchmark Multidimensionnel des Large Language Models (LLM)

Université Paris 8 Vincennes–Saint–Denis

Master Informatique — Ingénierie en IA (I2A)

Projet de fin de semestre

Étudiants :
Mohamed Moudir
Anas El Alaoui

Encadrant : Mahdi Ammi

Projet Master 1 I2A

Table des matières

Chapitre 1

Introduction

L'essor de l'intelligence artificielle générative a profondément transformé les usages numériques contemporains. Les *Large Language Models* (LLM) occupent une place centrale dans cette évolution en raison de leur capacité à comprendre, raisonner et produire du langage naturel avec un haut niveau de précision.

Dans le contexte universitaire, ces modèles ouvrent de nouvelles perspectives pour l'accompagnement pédagogique, le soutien à la recherche et l'automatisation de tâches administratives. Toutefois, leur intégration soulève des enjeux majeurs liés à la fiabilité des réponses, à l'expérience utilisateur, à l'impact environnemental et aux risques éthiques.

Ce projet propose une évaluation comparative et multidimensionnelle de plusieurs LLM afin d'identifier le modèle offrant le meilleur compromis pour une intégration responsable au sein d'une université francophone.

Chapitre 2

État de l'art

2.1 Fondements des Large Language Models

Les LLM reposent sur des architectures de type *Transformer*, introduites pour améliorer la modélisation des dépendances longues dans les textes. Entraînés sur des corpus massifs, ces modèles acquièrent des capacités avancées de généralisation, de raisonnement et de compréhension contextuelle.

2.2 Usages académiques des LLM

Les universités envisagent l'utilisation des LLM pour :

— l'assistance aux étudiants (tutorat intelligent, aide à la compréhension),
— le soutien aux enseignants (création de supports pédagogiques),
— l'automatisation de services administratifs.

Cependant, ces usages doivent être encadrés afin de garantir la neutralité académique et la qualité des contenus produits.

2.3 Limites des évaluations classiques

Les benchmarks traditionnels, tels que le MMLU, se concentrent principalement sur la performance cognitive. Or, ces évaluations sont insuffisantes pour juger de la pertinence d'un LLM dans un environnement réel, où des contraintes techniques, environnementales et éthiques doivent être prises en compte.

2.4 Responsabilité environnementale et éthique

Les travaux récents soulignent l'importance de mesurer l'empreinte carbone des systèmes d'IA et de contrôler les biais algorithmiques. Des outils spécialisés permettent aujourd'hui de compléter les évaluations cognitives par des analyses environnementales et éthiques.

Chapitre 3

Objectifs et problématique

La problématique centrale de ce projet est la suivante :

> *Quel modèle de langage offre le meilleur compromis entre performance cognitive, efficience opérationnelle, responsabilité environnementale et sécurité éthique pour une intégration universitaire ?*

Les objectifs sont :

— comparer plusieurs LLM selon des critères multidimensionnels,

— fournir des indicateurs concrets pour orienter les choix institutionnels,

— proposer une méthodologie reproductible. Ces objectifs s'inscrivent dans une démarche pragmatique visant à rapprocher les performances théoriques des modèles de leurs usages réels. En effet, un modèle performant sur un benchmark cognitif peut se révéler inadapté dans un contexte opérationnel s'il présente une latence excessive, un coût énergétique élevé ou des risques éthiques non maîtrisés.

  Ainsi, ce projet adopte une approche globale permettant d'éclairer les choix technologiques de l'Université à partir de critères mesurables, transparents et reproductibles.

Chapitre 4

Modèles évalués

Trois modèles de langage ont été sélectionnés afin de représenter différentes approches technologiques et stratégiques du domaine des LLM.

GPT-4o est utilisé comme modèle de référence. Il se distingue par ses capacités avancées de raisonnement, sa polyvalence et sa robustesse sur des tâches académiques complexes. Son intégration dans cette étude permet de disposer d'un point de comparaison représentatif de l'état de l'art industriel.

Mistral Large constitue une alternative européenne crédible. Il a été retenu pour évaluer les performances d'un modèle optimisé pour la langue française et pour répondre aux enjeux de souveraineté numérique, particulièrement importants dans un contexte institutionnel.

Llama-3.3 (via Groq) a été sélectionné afin d'analyser l'impact d'une infrastructure matérielle spécialisée sur les performances d'inférence. Ce modèle permet d'étudier le compromis entre rapidité, consommation énergétique et qualité des réponses.

Ce choix de modèles permet ainsi une comparaison équilibrée entre performance cognitive, efficacité opérationnelle et considérations stratégiques.

Chapitre 5

Données et méthodologie

5.1 Jeu de données

L'évaluation repose sur le benchmark **MMLU-FR**, couvrant 57 domaines académiques. Un échantillon de 100 questions a été sélectionné afin de garantir un compromis entre fiabilité statistique et faisabilité expérimentale.

5.2 Protocole expérimental

Le protocole comprend :

- des appels API asynchrones,
- une extraction automatisée des réponses,
- une analyse éthique via Toxic-BERT,
- une estimation de l'empreinte carbone avec CodeCarbon.

Chapitre 6

Indicateurs de performance

Les indicateurs de performance ont été définis afin de couvrir l'ensemble des dimensions pertinentes pour une intégration universitaire des LLM.

La **performance cognitive** est mesurée à l'aide du taux de réussite sur le benchmark MMLU-FR. Cet indicateur reflète la capacité du modèle à mobiliser des connaissances académiques variées et à fournir des réponses correctes.

La **latence moyenne** correspond au temps écoulé entre l'envoi de la requête et la réception de la réponse. Elle constitue un critère déterminant pour les usages interactifs, tels que les assistants pédagogiques.

L'**empreinte carbone** est estimée en grammes de CO_2 équivalent par requête. Cet indicateur permet d'évaluer la compatibilité des modèles avec les objectifs de sobriété numérique et de développement durable.

Enfin, la **sécurité éthique** est évaluée à travers les scores de toxicité et de biais identitaires. Ces métriques permettent de mesurer les risques de dérives langagières incompatibles avec les valeurs académiques.

Chapitre 7

Résultats expérimentaux

7.1 Résultats globaux

TABLE 7.1 – Résultats du benchmark multidimensionnel (N = 100)

Modèle	Précision	Latence (s)	CO_2 (g)	Toxicité	Biais
GPT-4o	83%	0.64	0.13	0.0012	0.00014
Llama-3.3 (Groq)	76%	1.56	0.50	0.0011	0.00014
Mistral Large	50%	0.31	0.20	0.0006	0.00008

7.2 Analyse détaillée des performances

Les résultats montrent que GPT-4o obtient la meilleure précision cognitive, confirmant sa robustesse sur des domaines académiques variés. Llama-3.3 se positionne comme une alternative performante, tandis que Mistral Large présente une précision plus faible sur cet échantillon.

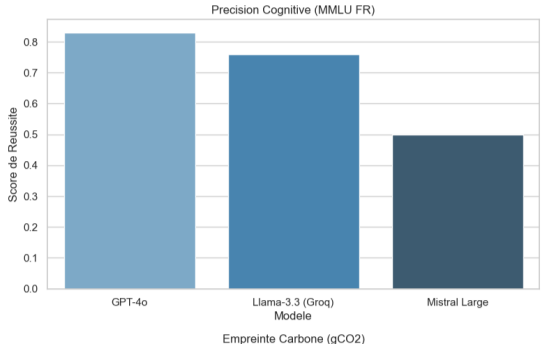
Sur le plan technique, Mistral Large se distingue par une latence très faible, ce qui constitue un avantage pour des applications interactives. En revanche, Llama-3.3 affiche la latence la plus élevée dans ce contexte expérimental.

D'un point de vue environnemental, GPT-4o présente l'empreinte carbone la plus faible, tandis que Llama-3.3 montre un impact plus important. Enfin, les scores éthiques sont très faibles pour l'ensemble des modèles, attestant de leur compatibilité avec un usage académique.

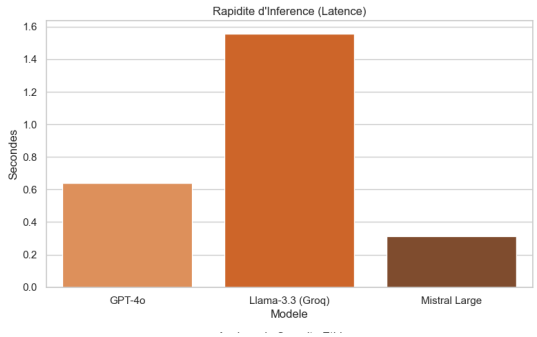
7.3 Éthique et sécurité

L'analyse éthique constitue un critère fondamental dans le cadre d'une intégration des modèles de langage en milieu universitaire. Elle vise à garantir la neutralité des contenus générés ainsi que l'absence de dérives telles que la toxicité ou les biais identitaires.

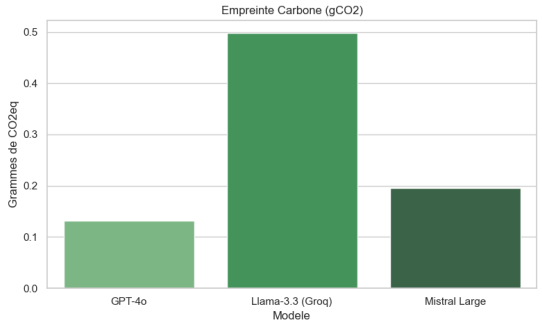
Afin d'illustrer les résultats de l'audit éthique, une représentation graphique détaillée a été réalisée à partir des scores fournis par le classificateur Toxic-BERT.



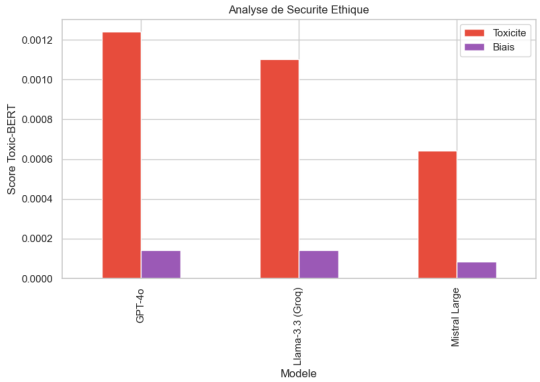
(a) Score de toxicité par modèle



(b) Score de biais identitaires par modèle



(c) Comparaison globale des indicateurs éthiques



(d) Vue synthétique de la sécurité éthique

FIGURE 7.1 – Analyse graphique des indicateurs de sécurité éthique (toxicité et biais) pour les modèles LLM évalués.

Chapitre 8

Limites de l'étude

Cette étude présente plusieurs limites :

- taille réduite de l'échantillon,
- dépendance aux infrastructures API,
- absence d'évaluation qualitative humaine,
- estimation environnementale dépendante du matériel local.

Chapitre 9

Perspectives et recommandations

Les perspectives incluent :

— une évaluation à plus grande échelle,

— l'étude de cas d'usage réels,

— la mise en place de politiques institutionnelles d'usage responsable,

— l'intégration de mécanismes de supervision éthique.

Chapitre 10

Conclusion générale

Ce rapport met en évidence l'intérêt d'une approche multidimensionnelle pour l'évaluation des LLM. GPT-4o apparaît comme la solution la plus équilibrée pour une intégration universitaire globale, tandis que Mistral Large et Llama-3.3 peuvent répondre à des besoins spécifiques selon les contraintes opérationnelles.

Ce travail fournit une base méthodologique solide pour accompagner la prise de décision stratégique des établissements d'enseignement supérieur.