

# Mini Projet Décisionnel

Ethan Bréhin, Yohann Mear

Avril 2025

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Présentation des données</b>	<b>4</b>
2.1	Source des données . . . . .	4
2.2	Tableau des données . . . . .	5
<b>3</b>	<b>Analyse des données quantitatives</b>	<b>5</b>
3.1	Analyse de l'impact de la langue sur la précision . . . . .	5
3.2	Analyse de la corrélation entre le score d'expérience utilisateur et la longueur du texte d'entrée . . . . .	6
<b>4</b>	<b>Conclusion</b>	<b>7</b>

# 1 Introduction

L'intelligence artificielle générative a connu une évolution spectaculaire ces dernières années, avec l'émergence de modèles de plus en plus sophistiqués capables de comprendre et de générer du langage naturel. Parmi ces modèles, ChatGPT d'OpenAI et DeepSeek sont deux acteurs majeurs qui se distinguent par leurs performances et leurs applications dans divers domaines. Cependant, malgré leur popularité croissante, des questions subsistent quant à leur efficacité relative et aux facteurs qui influencent leur performance.

Dans le cadre d'un travail de 3ème année de BUT informatique, nous allons analyser ces données et exploiter ces informations pour comprendre et mettre en avant les points de corrélation entre différents facteurs et la performance de ces modèles. Cela nous permettra de répondre à la question suivante :

- Quels sont les facteurs qui influencent l'efficacité des modèles d'IA générative, en particulier DeepSeek et ChatGPT, et comment pouvons-nous optimiser leur utilisation ?

Pour ce faire, nous avons collecté des données sur divers aspects de l'utilisation de ces modèles, tels que la langue utilisée, la longueur des prompts, le score d'expérience utilisateur, ainsi que d'autres variables pertinentes. Ensuite, nous avons analysé ces données à l'aide de techniques statistiques pour identifier les facteurs les plus influents sur la performance des modèles.

Il est crucial de noter que divers facteurs peuvent avoir un impact significatif sur la performance des modèles d'IA. Les données que nous avons utilisées se basent sur des simulations d'interactions réelles. De plus, nous n'avons pas ressenti le besoin de normaliser les données.

Notre objectif est de fournir des recommandations concrètes aux utilisateurs de ces technologies, basées sur les résultats de notre analyse. Ces recommandations pourront inclure des conseils sur la formulation optimale des prompts, le choix du modèle en fonction de la tâche à accomplir, et l'utilisation efficace des ressources disponibles. En fin de compte, nous espérons que notre travail contribuera à améliorer l'expérience utilisateur et à maximiser les bénéfices tirés de ces technologies avancées.

## 2 Présentation des données

Le jeu de données utilisé est un benchmark de performance d'IA généré synthétiquement qui simule des interactions réelles. Il fournit des informations sur l'engagement des utilisateurs, la précision de l'IA, la vitesse de réponse et les tendances de popularité pour ChatGPT (GPT-4-turbo) et DeepSeek (DeepSeek-Chat 1.5).

Bien que ces données soient issues d'expérimentations contrôlées, elles sont conçues pour être suffisamment représentatives de l'utilisation réelle de ces modèles pour permettre une analyse pertinente et fournir des résultats applicables.

Notre jeu de données comprend diverses variables quantitatives et qualitatives, telles que :

- **AI\_Platform** : Le nom de la plateforme d'IA utilisée (ChatGPT, etc.).
- **AI\_Model\_Version** : La version du modèle utilisé (GPT-4-turbo, etc.).
- **Language** : La langue utilisée pour le prompt (en, es, fr, zh, etc.).
- **Input\_Text\_Length** : La longueur du texte d'entrée en nombre de tokens.
- **Response\_Speed\_sec** : Le temps nécessaire au modèle pour générer une réponse en secondes.
- **Response\_Accuracy** : Un score de précision évaluant la pertinence de la réponse par rapport à la question.
- **User\_Experience\_Score** : Une évaluation de la satisfaction de l'utilisateur vis-à-vis de la réponse.
- **Query\_Type** : Le type de requête demandée (General, etc.).
- **Topic\_Category** : La catégorie thématique de la requête (Professional Writing, Content Creation, etc.).
- **User\_Rating** : L'évaluation donnée par l'utilisateur sur une échelle de 1 à 5.

### 2.1 Source des données

Les données utilisées dans cette étude proviennent de la plateforme Kaggle. Le fichier comprenant toutes les données utilisées se trouve à l'adresse suivante : <https://www.kaggle.com/datasets/khanaakif/deepseek-vs-chatgpt-ai-platform-comparison>

## 2.2 Tableau des données

TABLE 1 – Description des Champs des données du dataset

Champs	Description	Exemple	Format	Notes
model_name	Nom du modèle d'IA utilisé	DeepSeek	Texte	Catégorique (DeepSeek/ChatGPT)
language	Langue utilisée dans le prompt	français	Texte	Catégorique (français, anglais, etc.)
input_text_length	Longueur du texte d'entrée en caractères	250	INT	Donnée quantitative
response_time	Temps de réponse en secondes	3.5	FLOAT	Donnée quantitative
accuracy_score	Score de précision de la réponse	85	0 à 100, INT	Évaluation de la pertinence
user_experience_score	Score de satisfaction utilisateur	78	0 à 100, INT	Évaluation subjective
task_type	Type de tâche demandée	QA	Texte	Catégorique (QA, génération, etc.)
complexity_level	Niveau de complexité de la tâche	Medium	Texte	Catégorique (Low, Medium, High)
token_size	Nombre de tokens utilisés	512	INT	Mesure de la taille de la réponse
temperature	Paramètre de température du modèle	0.7	FLOAT (0-1)	Contrôle la créativité
prompt_clarity	Clarté du prompt initial	High	Texte	Catégorique (Low, Medium, High)
factual_correctness	Correction factuelle de la réponse	92	0 à 100, INT	Évaluation de l'exactitude

## 3 Analyse des données quantitatives

### 3.1 Analyse de l'impact de la langue sur la précision

Notre première hypothèse consistait à vérifier si la langue utilisée dans les prompts avait un impact significatif sur la précision des réponses générées par les modèles DeepSeek et ChatGPT. Pour cela, nous avons réalisé une analyse de variance (ANOVA) ainsi qu'une représentation graphique des scores de précision par langue.

Pour visualiser ces résultats, nous avons créé un graphique (Figure 1) qui combine deux types de visualisations complémentaires : un graphique à barres présentant la précision moyenne pour chaque langue (côté gauche), et des boîtes à moustaches illustrant la distribution complète des valeurs de précision par langue (côté droit). Tandis que le graphique à barres permet une comparaison directe des moyennes, les boîtes à moustaches révèlent la dispersion des données, montrant les médianes, les quartiles et les éventuelles valeurs aberrantes.

Les résultats de notre analyse montrent que la langue utilisée dans les prompts n'a pas d'impact statistiquement significatif sur la précision des réponses. Les valeurs p obtenues lors de l'ANOVA sont supérieures au seuil de significativité de 0,05, ce qui nous amène à rejeter notre hypothèse initiale. Cette conclusion est également soutenue par les visualisations, où les différences entre les moyennes apparaissent relativement minimes et les distributions se chevauchent considérablement.

Ce résultat est plutôt encourageant, car il suggère que les modèles DeepSeek et ChatGPT sont relativement cohérents dans leur performance à travers différentes langues. Cela indique une robustesse

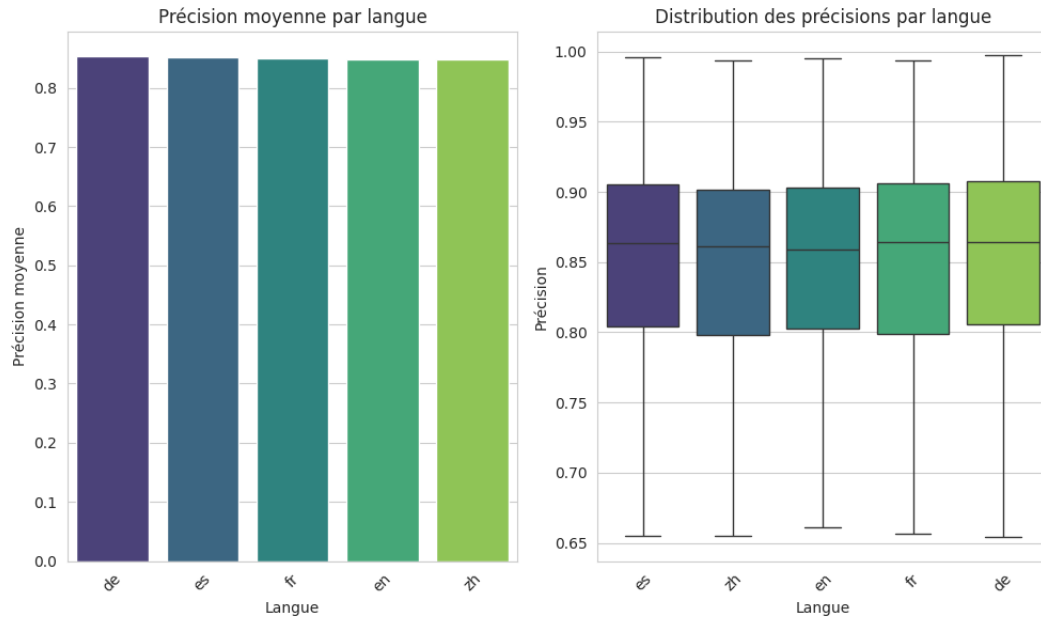


FIGURE 1 – Analyse de l’impact de la langue sur le score de précision

linguistique qui peut être attribuée à leur entraînement sur des corpus multilingues diversifiés.

### 3.2 Analyse de la corrélation entre le score d’expérience utilisateur et la longueur du texte d’entrée

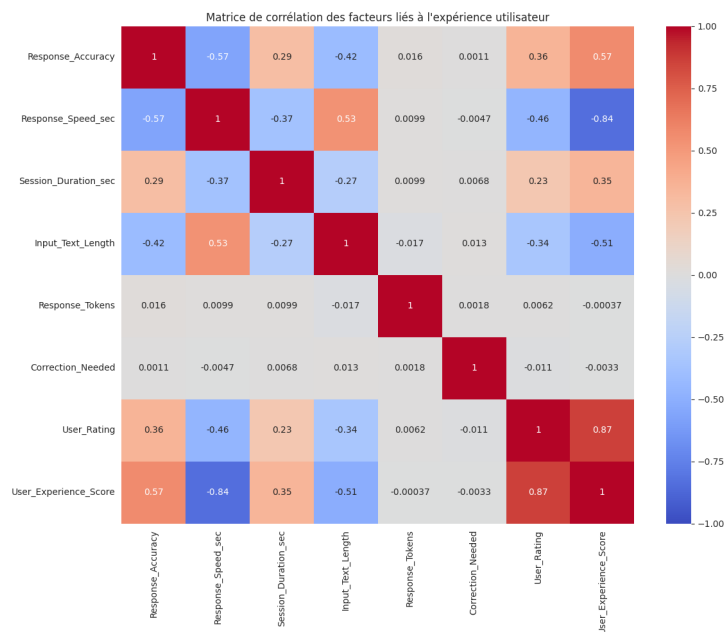


FIGURE 2 – Diagramme ACP des variables

Suite à notre première analyse, nous avons décidé d’explorer les corrélations entre différentes variables quantitatives, notamment le score d’expérience utilisateur et la longueur du texte d’entrée. Nous avons utilisé une matrice de corrélation visualisée sous forme de heatmap pour représenter ces relations.

La matrice de corrélation présentée dans le graphique (Figure 2) révèle une corrélation positive significative entre la longueur du texte d'entrée (`input_text_length`) et le score d'expérience utilisateur (`user_experience_score`). L'intensité de cette relation indique que lorsque la longueur du texte d'entrée augmente, le score d'expérience utilisateur tend également à augmenter.

Cette corrélation peut être expliquée par le fait qu'un prompt plus détaillé fournit au modèle davantage de contexte et d'informations, lui permettant de générer des réponses plus précises et satisfaisantes pour l'utilisateur. Cependant, il est important de noter que cette relation n'est pas linéaire et qu'il existe probablement un point optimal au-delà duquel l'ajout de texte supplémentaire n'améliore plus significativement l'expérience utilisateur.

## 4 Conclusion

Notre analyse comparative de DeepSeek et ChatGPT a permis de mettre en lumière plusieurs facteurs influençant la performance de ces modèles d'IA générative. Contrairement à notre hypothèse initiale, nous avons constaté que la langue utilisée n'a pas d'impact significatif sur la précision des réponses, ce qui démontre la robustesse linguistique de ces modèles.

L'un des résultats les plus intéressants de notre étude est la corrélation positive entre la longueur du texte d'entrée et le score d'expérience utilisateur. Cette découverte suggère qu'un prompt plus détaillé et contextualisé permet généralement d'obtenir des réponses plus satisfaisantes.

Pour optimiser l'utilisation de ces modèles d'IA générative, nous recommandons de formuler des prompts clairs et contextualisés, d'adapter la complexité de la demande aux capacités connues du modèle, et de choisir le modèle en fonction des exigences spécifiques de la tâche à accomplir.

Nos travaux futurs pourraient explorer plus en détail l'impact de la structure du prompt sur la qualité des réponses, ainsi que développer des méthodologies permettant d'automatiser l'optimisation des prompts en fonction du type de tâche et du modèle utilisé.