



MÉMOIRE DE MASTER 1

SORBONNE UNIVERSITÉ
MASTER DE SCIENCES DU LANGAGE

PARCOURS LANGUE ET INFORMATIQUE

Alexis Loubes et Valérie Peng

Sous la direction de : Laurence Devillers

LES INTERACTIONS HUMAIN-MACHINE DANS L'ÉVALUATION

Année universitaire : 2023-2024

REMERCIEMENTS

Nous tenons à exprimer toute notre reconnaissance et notre gratitude envers tous ceux qui nous ont donné les moyens de mener à bien ce mémoire.

Nous remercions notre encadrante Madame Laurence DEVILLERS ainsi que Monsieur Claude MONTACIÉ pour leur patience , leur gentillesse ainsi que tout le temps qu'ils nous ont consacré pour ce travail. Leur expertise nous a été d'une grande aide tout au long de la réalisation de ce mémoire.

De plus , nous voulons remercier notre entourage ainsi que toutes les personnes qui ont contribué de près ou de loin à l'aboutissement et au bon déroulement de ce travail.

Enfin , nous exprimons également nos remerciements au sein du binôme , nous tenons à nous remercier mutuellement pour l'entraide , la coopération et l'encouragement que nous nous sommes données mutuellement.

RÉSUMÉ

“Si l’humain est aussi biaisé que la machine, ne faudrait-il pas les évaluer ensembles ?”

C’est à travers une société qui intègre de plus en plus les grands modèles de génération textuelle à ses activités que les évaluations tentent de mesurer et de tracer les frontières de ce que ces outils peuvent faire ou non.

Nous montrons dans ce mémoire que lorsque 62 étudiants universitaires font face à différentes tâches en situation d’examen, ils n’utilisent pas ChatGPT de la même façon et n’aboutissent pas aux mêmes résultats en présence d’un même outil.

Un outil dont les performances sont chiffrées dans les benchmarks et qui peuvent être considérées à tort comme fixes et absolues.

SOMMAIRE

RÉSUMÉ.....	ii
SOMMAIRE.....	2
LISTE DES ABRÉVIATIONS.....	2
1. INTRODUCTION.....	3
1.1 <i>L'arrivée de ChatGPT.....</i>	<i>3</i>
1.2 <i>Focus sur ChatGPT.....</i>	<i>4</i>
1.3 <i>Les modèles de langages.....</i>	<i>5</i>
1.4 <i>Un chemin jusqu'aux LLM.....</i>	<i>6</i>
1.5 <i>Variabilité du LLM.....</i>	<i>8</i>
1.6 <i>Défis.....</i>	<i>10</i>
1.7 <i>Enjeux.....</i>	<i>12</i>
1.8 <i>Problématique.....</i>	<i>13</i>
2. ÉTAT DE L'ART.....	14
2.1 <i>Les benchmarks : Un outil d'évaluation.....</i>	<i>14</i>
2.2 <i>Une étude complète.....</i>	<i>17</i>
2.3 <i>Une étude centrée sur les humains.....</i>	<i>20</i>
3. LE CORPUS.....	22
3.1 <i>But du corpus.....</i>	<i>22</i>
3.2 <i>Contenu du corpus.....</i>	<i>22</i>
3.3 <i>Description des tâches du corpus.....</i>	<i>23</i>
3.4 <i>Méthodologie d'exploitation du corpus.....</i>	<i>23</i>
3.5 <i>Les métriques d'évaluation.....</i>	<i>24</i>
4. L'EXPÉRIENCE.....	26
4.1 <i>Contexte de l'expérience.....</i>	<i>26</i>
4.2 <i>Le choix de l'outil.....</i>	<i>26</i>
4.3 <i>Déroulement de l'épreuve.....</i>	<i>29</i>
5. RÉSULTATS.....	30
5.1 <i>La réussite : ChatGPT améliore-t-il les résultats ?.....</i>	<i>30</i>
5.2 <i>Les questions.....</i>	<i>33</i>
5.3 <i>Les étudiants.....</i>	<i>36</i>
5.4 <i>L'interaction quantitative.....</i>	<i>38</i>
5.5 <i>L'interaction qualitative.....</i>	<i>41</i>

5.6 <i>Discussion et limites</i>	43
6. CONCLUSION	44
TABLEAU DES ANNEXES	45
TABLE DES FIGURES	51
BIBLIOGRAPHIE	52

LISTE DES ABRÉVIATIONS

ChatGPT : Chat Generative Pre-trained Transformer

NLP : Natural Language Processing (Traitement du Langage Naturel)

IA : Artificial Intelligence (Intelligence artificielle)

LLM : Large Language Model

HMM : Hidden Markov Models (Modèles de Markov Cachées)

RNN : Réseaux de neurones récurrents

BERT : Bidirectional Encoder Representations from Transformer

MMLU : Massive Multi-Task Language Understanding (Compréhension du Langage à Grande Échelle et Multi-Tâches)

GSM8K : Grade School Math 8K (Mathématiques de l'École Primaire 8K)

ARC Challenge : Abstraction and reasoning challenge

MGSM : Multilingual Grade School Math

MINT : multi-turn interactions

BLEU : Bilingual Evaluation Understudy

Llama : Large Language Model Meta AI (Modèle de Langage à Grande Échelle de Meta AI)

1. INTRODUCTION

1.1 L'arrivée de ChatGPT

Depuis l'arrivée de **ChatGPT**, l'utilisation de l'IA s'est ancrée dans le quotidien de beaucoup de personnes. Le site sur lequel se trouve l'outil n'a pris que 5 jours avant d'atteindre le **million d'utilisateurs** après son lancement en novembre 2022 et a reçu plus de 600 millions de visites durant le mois de janvier qui a suivi ce lancement. [[FD 2024](#)]

Malgré les chiffres, il faut rappeler que l'accès à internet est inégal selon les pays et que les considérations précédentes sont à modérer : on ne parle pas des humains en général mais de ceux qui ont un accès au numérique avec internet ce qui n'est pas aussi universel que l'on pense. Selon l'ITU environ 67% de la population mondiale a accès à internet. [[ITU 2022](#)]

Le modèle de langage ChatGPT a été créé par OpenAI. Le GPT provient du terme « Generative Pre-trained Transformer ».

Lorsqu'il est interrogé sur ses compétences, il répond qu'il est capable de générer du texte afin d'aider l'utilisateur dans différentes tâches liées au langage naturel, telles que la création de texte, la compréhension du langage, la traduction ou même la réalisation de résumés.

Cela correspond à des **problématiques du TAL** (traitement automatique des langues) également appelées NLP (natural language processing).

L'outil semble être en voie d'intégration dans différents domaines comme :

- **Les entreprises** : marketing et ventes, opérations, informatique/ingénierie, risque et juridique, ressources humaines, comptabilité et finance ou optimisation des services publics/employés.
- **La santé** : interaction avec les patients, soutien au diagnostic clinique, services de télésanté, éducation en santé.
- **Création de contenu** : L'industrie du journalisme.

Il est donc intéressant de se pencher sur ce qu'est vraiment cet outil très polyvalent en apparence.

1.2 Focus sur ChatGPT

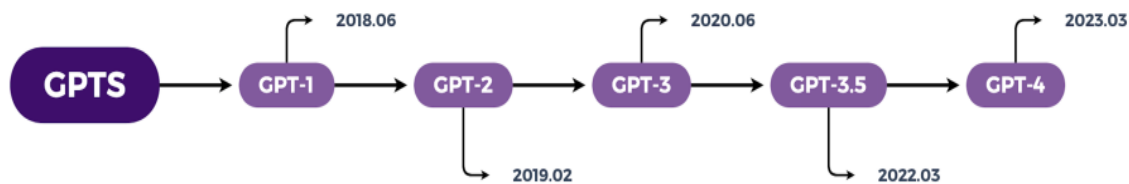


Figure 1 - Frise chronologique de l'évolution des modèles GPT

ChatGPT est un modèle de langage dont la première version a été mise à disposition du grand public en novembre 2022 par l'entreprise OpenAI. Ce modèle de langage regroupe plusieurs versions différentes montrées dans la *figure 1*. [[OSSEFG 2023](#)]

La dernière version mise à disposition à ce jour est GPT-4 sortie en mars 2023.

L'acronyme de GPT signifie Generative Pre-trained Transformer et montre que ce modèle fait recours aux **transformers**. Les transformers désignent une architecture de modèle se basant sur des mécanismes antérieurement existants avec de nouveaux mécanismes innovants qui sont responsables du paysage actuel des IA de génération textuelle. [[AIAYN 2023](#)]

Dans une enquête menée en mars 2024 par l'**IFOP**, sur un échantillon d'environ 1000 personnes (qui se dit représentatif de la population française), 25% déclarent avoir eu recours à une IA générative. ChatGPT représente 66% des IA génératives citées. Sur ce quart, un tiers l'utilise dans sa vie professionnelle. [[IFOP 2024](#)]

Cette présence croissante peut inquiéter sur l'avenir des métiers. En effet le rapport de la banque d'investissement Goldman Sachs affirme qu'en moyenne 18% de l'activité professionnelle pourrait être automatisée par l'IA dont en grande partie l'IA générative. [[GS 2023](#)]

Il faut cependant préciser plusieurs choses : nous avons parlé d'un paysage des IA génératives. Cela veut donc dire d'une part que ChatGPT ne compose pas ce paysage à lui seul et d'autres parts que cet ensemble de modèles est issu d'une agrégation d'innovations et de travaux au fil du temps.

1.3 Les modèles de langages

Un modèle de langage est défini dans le GDT (Grand dictionnaire terminologique) comme un *“Modèle statistique permettant, à partir de données sur la distribution d'éléments dans une langue naturelle, de déterminer pour chacun d'eux la possibilité qu'il apparaisse comme prochain élément d'une chaîne donnée.”* [[GDT 2023](#)]

Cette définition est assez moderne puisqu'elle exclut les premiers systèmes à base de règles.

Le fonctionnement des premiers modèles statistiques simples est principalement basé sur des **statistiques** établies avec un **corpus** textuel. Son but est de prédire la suite la plus probable/fréquente d'une séquence de mots donnée.

Au vu du but d'un modèle de langage, peut-on parler d'intelligence artificielle ?

- Selon **Marvin Lee Minsky**, une intelligence artificielle est *“ la construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau...”* [[WIKI 2023](#)]
- Dans le **Larousse**, l'IA est décrite comme un *“Ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine.”* [[DLR IA](#)]

Selon ces définitions et si nous admettons que les modèles comme ChatGPT traitent effectivement les tâches du TAL -Des tâches par ailleurs que certains outils effectuent déjà sans être considérés en tant qu' IA- alors nous pouvons les considérer comme des IA, dans le cas où ils les exécutent de façon satisfaisante bien sûr.

Prenons en considération un bref historique des modèles de langage pour comprendre ce qui différencie les premiers modèles de ceux du type de ChatGPT qui sont **les grands modèles de langage** également appelés **LLM** (Large language model). Ici l'abréviation **LLM** désignera par abus de langage les grands modèles de génération textuelle qui sont une catégorie de LLM.

1.4 Un chemin jusqu'aux LLM

Tout d'abord, les principales avancées ont été influencées en partie par des progrès techniques sur les ordinateurs engendrant une augmentation exponentielle de la puissance de calcul. De plus, ces avancées se sont influencées entre elles et certaines ont même rendu de nouveau populaire des théories considérées jusqu'alors comme dépassées.

Ci-dessous figure une liste non exhaustive d'importantes évolutions dans le domaine par périodes approximatives:

- **1950-1970 - Les premiers modèles à base de règles**

Joseph Weizenbaum conçoit ELIZA en 1966, un programme informatique simulant une conversation grâce à des règles simples et des correspondances de motifs. Les réponses étaient générées avec des script prédéfinies et elles étaient déclenchées par des mots-clés. [[CL 1966](#)]

- **1980-1990 - Les modèles statistiques : le modèle n-gram**

Le modèle n-gram va calculer la probabilité qu'un mot apparaisse après une séquence contenant un nombre n de mots. Il prédit grâce à des fréquences d'apparition. Par exemple, quand n=2 (modèle bigramme), si la fréquence de "suis" est plus élevée que "est" après "Je" alors sa probabilité d'apparaître sera elle aussi plus élevée et le modèle choisira l'élément le plus probable dans sa prédiction du mot suivant. [[SHANNON 1948](#)]

- **1970-2000 - De nouveaux modèles statistiques :**

- Modèles de Mélange Gaussien et Modèles Conditionnels [[CRF 2001](#)]
- Modèles de Markov Cachés (HMM) [[IEEE 1986](#)]

- Modèles à Base de Maximum d'Entropie [[CORE 1997](#)]

- **1985 - L'Algorithme de rétropropagation pour les réseaux de neurones**

Cet algorithme a été présenté par Ronald J. Williams, David Rumelhart et Geoffrey Hinton. Il révolutionne les **réseaux de neurones** en optimisant les poids dans les réseaux multicouches et résolvant des problèmes non linéaires que le perceptron multicouche simple (MLP) peine à traiter. [[LIRER 1985](#)]

- **1998 - Les réseaux de neurones à mémoire à long terme (LSTM)**

Sepp Hochreiter et Jürgen Schmidhuber développent une amélioration des réseaux de neurones récurrents (**RNN**) qui se révèle efficace pour une variété de tâches du TAL comme la traduction automatique, la reconnaissance de la parole, et l'analyse de sentiment. Le terme paramètre qui sera utilisé, désigne les poids et biais dont le modèle se sert pour des calculs qui se font quand les informations passent dans les neurones. On va souvent associer la complexité d'un modèle à son **nombre de paramètres**. [[LSTM 1997](#)]

- **2013 - Word2Vec**

Google sort **Word2Vec** en 2013. Le mécanisme important à retenir est une nouvelle représentation des mots en vecteurs de dimensions fixes égales à la taille du vocabulaire captant le contexte et les relations sémantiques. Les applications directes sont la prédiction d'un mot selon le contexte autour et l'inverse.

Il popularise la représentation des mots en vecteur : **les word embeddings (plongement lexicaux en français)** qui sont toujours utilisés dans les modèles récents. [[MIKOLOV 2013](#)]

- **2017 - L'arrivée des Transformers**

L'architecture des Transformers utilise entre autres un **mécanisme d'attention** plutôt que des réseaux de neurones récurrents. Ce mécanisme permet au modèle de se concentrer sur l'information importante tout en considérant l'ensemble du contexte simultanément. Le modèle va en sortie donner un vecteur de même taille que le vecteur en entrée avec des probabilités pour chaque mot dans celui-ci.

On trouve une couche de Transformers dans la plupart des modèles actuellement populaires comme **GPT, Llama et Bard**. [[AIAYN 2023](#)]

- Depuis 2018 - Les premiers modèles larges (LLM) :

C'est avec **BERT** possédant plus de 100 millions de paramètres par Google en 2018 que l'ère des grands modèles s'enclenche. En 2019, **GPT-2** de OpenAI arrive avec 1,5 milliard de paramètres puis en 2021, **GPT-3** avec ses 175 milliards de paramètres. Et depuis, de nouveaux modèles continuent à voir le jour avec des tailles de plus en plus vertigineuses.

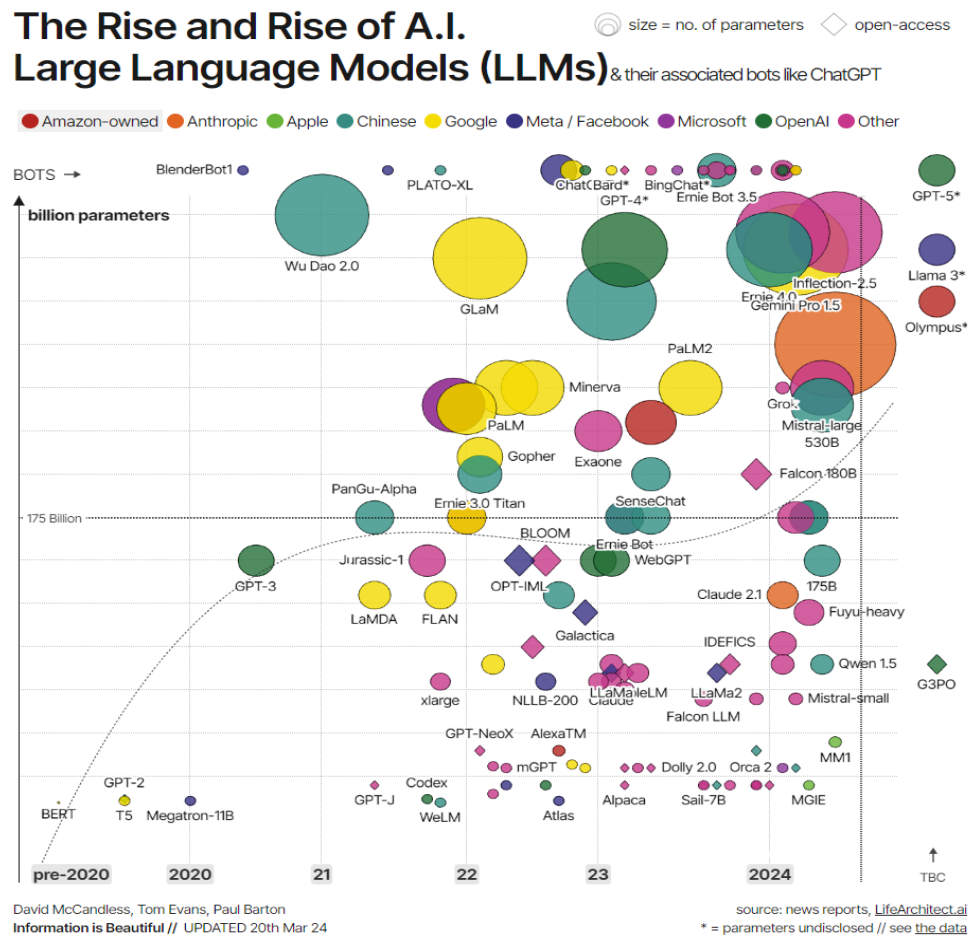


Figure 2 - Graphique représentatif de l'augmentation des nouveaux LLM

De cette explosion des LLM et de leur utilisation croissante dans différents domaines, découlent naturellement des défis. Des défis qu'on peut souvent associer à des risques.

1.5 Variabilité du LLM

En premier lieu, il est important de faire la différence entre la sortie/réponse (**l'output**) et la tâche que le LLM a effectué. Bien qu'à chaque fois la sortie soit textuelle, la tâche peut

relever de : la classification, l'annotation, un résumé, une réponse à des questions diverses de connaissance ou des propositions d'idées.

En réalité, il est assez difficile d'établir des frontières sur les tâches qu'un LLM peut faire ou non tant elles peuvent sembler diverses.

D'autant plus que plusieurs paramètres peuvent faire varier la qualité de la réponse. Il faut les prendre en considération.

- *Le modèle est-il ajusté où non pour ce genre de tâches ?*

C'est le **fine tuning**, une technique qui va entraîner une seconde fois un modèle sur d'autres données pour le spécialiser sur des tâches (GPT est le modèle de base et ChatGPT une version "ajustée" pour l'interaction avec l'humain). [\[FLM 2022\]](#)

- *Est-ce que le texte d'entrée est correctement formé pour réaliser la tâche ?*

Ce texte s'appelle le **prompt**. On parle même de stratégie de prompting pour avoir les meilleures réponses possibles avec notre requête, accompagnée parfois d'un vocabulaire spécial qu'on va insérer dans le prompt pour aider le modèle (**tokens spéciaux**).

- *Comment est-ce que le modèle récupère les informations du prompt ?*

La récupération des informations du texte est influencée par la segmentation (**tokenisation**) du texte en unités plus petites appelées **tokens** et par les normalisations qu'ils peuvent subir. Ces tokens doivent être pertinents et correspondre aux tokens du **vocabulaire** préalablement établi pour le modèle, c'est-à-dire l'ensemble des unités qu'il accepte.

- *Le modèle choisi t-il la réponse la plus probable ou se permet-il d'être plus créatif ?*

C'est la **température**, plus elle est haute et plus le modèle va considérer des sorties moins probables. On dit qu'il va être moins déterministe dans ses choix.

- *Le modèle a-t-il pris en compte les échanges précédents pour contextualiser et améliorer sa réponse ?*

La gestion de la **mémoire** est importante puisque dans bien des cas certaines informations importantes se trouvent uniquement sur des prompts écrits auparavant. La version de base et “instruct” des modèles ne le font généralement pas mais la version “chat” est prévue pour.

1.6 Défis

Les premiers défis des LLM qui nous viennent partent de leur architecture gourmande

- Des besoins élevés en **jeux de données (datasets)**

Les LLM ont besoin de larges datasets pour apprendre. Des datasets qui sont peu présents dans beaucoup de domaines/langues et qui, par leur taille, deviennent dur à filtrer et analyser.

Pourtant ce qui se trouve sur le web source principale des datasets n’est pas toujours formé de contenus appréciables et représentatifs de tout ce que les gens peuvent dire ou penser (la science étant un discours avant tout). Ces données peuvent également devenir obsolètes ou trop anciennes pour un modèle dont on exige souvent une connaissance du monde à jour.

- Des besoins élevés en **puissance de calcul**

Malgré l’utilisation inédite des processeurs graphiques originalement dédiés aux jeux vidéos, l’apprentissage initial et le fine tuning restent gourmands en ressources utilisées faisant tourner pendant des jours entiers les ordinateurs. Ce qui implique également des coûts financiers, surtout si nous voulons souvent mettre à jour les données et donc relancer un énième apprentissage.

**Les seconds, nous parvenant dans l’exercice des machines, constituent des biais.
Mais qu’est ce qu’un biais ?**

“Biais” est un mot ayant plusieurs sens proches qui peuvent porter à confusion. Ici nous considérerons une définition proche des biais cognitifs. [\[WIKI 2024\]](#)

“Un biais peut être une déviation par rapport à une norme ou à une objectivité”

Une objectivité qu’on associe souvent à tort ou à raison à la vérité. Alors intuitivement les biais sont à éviter.

Il y a de très nombreux biais décrits de façons différentes, certains se chevauchent ou se ressemblent beaucoup. En général le nom du biais est lié à la raison de la déviation. Nous citerons ici quelques biais que nous trouvons pertinents pour les LLM.

- **Les biais algorithmiques : Les données que les humains récoltent**

Les distorsions systémiques qui prennent leurs racines dans le modèle. [\[DB 2024\]](#)

- Le biais de représentation : Comme dit plus haut, les données récoltées ne sont pas représentatives de ce que la population peut dire ou penser. Le résultat sera donc susceptible de ne pas être désirable.
- Le biais de préjugé : Le modèle est susceptible de reproduire les discriminations et stéréotypes présents sur internet et donc dans le dataset. Un dataset trop grand pour être correctement filtré. Le modèle choisissant le plus probable va donc amplifier le stéréotype

- **Les biais de l’humain en interaction avec la machine** [\[IQO 2024\]](#)

- Le biais d’automatisation : Le fait d’accorder une confiance excessive au modèle en inhibant tout jugement critique même si la réponse est fausse.
- Le biais de confirmation : interprétation des résultats du modèle d’une façon qui confirme nos croyances préexistantes.

Les biais entravent l’accès des LLM à un idéal d’objectivité et de performance. Ce sont

des biais venant des humains qui détériorent drastiquement les résultats pour des humains. En aucun cas les concepteurs ne peuvent imputer la responsabilité aux LLM s'il y a des effets néfastes. Ils sont conditionnés par des règles et des données constituées par leurs concepteurs..

Pourquoi parler d'effets néfastes ? Ont-ils réellement un impact plus étendu qu'une simple mauvaise réponse ?

1.7 Enjeux

Lors d'une utilisation domestique, une mauvaise réponse du LLM n'entraîne pas de conséquences importantes. Mais son usage dans le cadre d'un métier qui véhicule des enjeux rend l'exploitation de l'outil elle-même porteuse de risques.

De nombreuses études sont toujours en cours quant aux usages possibles des LLM mais nous savons déjà que des personnes les utilisent en entreprise ou songent à en déployer.

Si notre utilisation du LLM est biaisée, comment cela pourrait-il se répercuter si nous extrapolions ?

- Avocat: La fabulation.
- Médecine : Un diagnostic faux causant un danger potentiel.
- Recrutement : La reproduction de préjugés et discriminations multiples lors du choix.
- Finance : Une perte d'argent pouvant mettre à mal une entreprise.
- Création de contenu : Le vol de d'idées et de contenu sans citer l'auteur.
- Education : Perpétuation et transmission de préjugés ou de valeurs non désirables.

Les articles élogieux attisent cette popularité et ne cessent d'attribuer de nouvelles compétences aux LLM sans études concrètes.

Il faut donc évaluer ces LLM pour mesurer à quel point ils sont efficaces et leurs frontières pour éviter des utilisations totalement impertinentes aux capacités du LLM.

1.8 Problématique

Des évaluations existent déjà pour ces LLM mais elles ne semblent pas mettre l'humain au cœur de celles-ci. Des humains qui sont pourtant les principaux utilisateurs des LLM.

Si l'humain est aussi biaisé que la machine, ne faudrait-il pas les évaluer ensemble ?

Parvenir à montrer que les humains en présence d'un même outil de génération textuelle et en contexte d'utilisation concrète ont des résultats significativement variés remettrait en cause certains jeux de données d'évaluations et études.

Des jeux de données et études qui, dans ce cas, ne reflètent pas aussi bien les performances concrètes de ces modèles.

C'est pour cela que nous présenterons, après un état de l'art des évaluations qui nous ont inspiré, notre corpus élaboré via une expérience que nous avons menée.

Une expérience dont l'exploitation des résultats pourraient nous dire :

Dans quelles mesures est-il pertinent de mettre l'humain au centre de l'évaluation des grands modèles de génération textuelle ?

2. ÉTAT DE L'ART

Nous verrons les méthodes d'évaluations qui existent pour plusieurs LLM en s'attardant sur certaines pour mieux introduire notre propre évaluation.

2.1 Les benchmarks : Un outil d'évaluation

Le premier outil d'évaluation qui nous vient chronologiquement est celui du test de Turing. [\[TR 1950\]](#)

Pour valider ce test il faut que la personne ne puisse pas faire la différence lors d'un dialogue entre la machine et un humain. De nos jours, certains LLM passent certaines versions du test de Turing. Le problème de ce test est tout d'abord sa binarité : On le réussit ou non. De plus, la qualité du mimétisme de la machine vers l'humain n'est pas nécessairement preuve de son efficacité pour faire certaines tâches. [\[GPT_2023\]](#)

L'explosion du nombre de LLM a accéléré l'apparition de nouvelles évaluations pour comparer les LLM mais aussi pour mieux les vendre.

On appelle ces évaluations des **benchmarks** : Un ensemble de tâches standardisées de taille variable évaluant un ou plusieurs types de tâches. En général un benchmark regroupe plusieurs jeux de données fournissant des tâches/problèmes/consignes (avec la réponse dedans pour la correction) regroupés par thèmes.

En voici quelque exemples :

- **Général** : MMLU - Massive Multi-Task Language Understanding

“ test pour mesurer la précision multitâche d'un modèle de texte. Le test couvre 57 tâches, y compris les mathématiques élémentaires, l'histoire des États-Unis, l'informatique, le droit, et plus.” [\[MMMLU 2021\]](#)

- **Format** : Ce sont **15908** Questions en QCM des tâches dont les experts pour chaque domaine réussissent autour des **90%** de score tandis qu'une personne non experte a **34,5%** en moyenne.
- **Niveaux** : Ils sont variés, de l'école primaire jusqu'à l'université par exemple.
- **But** : Il est de mesurer à quel point le modèle peut extraire de la connaissance de son corpus massif pour résoudre un problème jusqu'à un niveau expert. Extraire les bonnes informations de son corpus par rapport au problème, on veut de lui une connaissance étendue du monde.

- **Mathématique** : GSM8K - Grade School Math

“ Pour diagnostiquer les échecs des modèles actuels et soutenir la recherche, nous présentons GSM8K, un ensemble de données de 8,5K problèmes de mots mathématiques de haute qualité et linguistiquement diversifiés de l'école primaire.” [[GSM8K 2021](#)]

- **Format** : Ce sont des problèmes mathématiques en langage naturel.
- **Niveaux** : Équivalent à l'école primaire donc un bon élève de collège est censé tout réussir.
- **But** : Il est de tester les capacités de raisonnement mathématique des modèles de langage en se concentrant sur des problèmes de mathématiques de niveau école primaire.
- **Autre** : L'évaluation prévoit une partie apprentissage avec un nombre de cycles pour que le modèle ait quelque exemple pour pouvoir répondre. Les réglages de l'entraînement sont faits avec GPT donc favorise ce modèle.

- **Raisonnement** : ARC Challenge - Abstraction and reasoning challenge

“Nous présentons un nouvel ensemble de questions, un corpus de textes et des lignes de base de référence assemblés [...] L'ensemble de ces éléments constitue le défi AI2 Reasoning Challenge (ARC), qui exige des connaissances et un raisonnement beaucoup plus puissants que les défis précédents tels que SQuAD ou SNLI.” [[ARC 2018](#)]

- **Format** : Il est en QCM avec **7787** questions séparées en entraînement et test.
- **Niveaux** : Il varie autour du niveau collège.

- **But** : Il s'agit de fournir des questions que les algorithmes classiques de récupération d'informations ne sont pas capables de résoudre.

- **Multilinguisme** : MGSM - Multilingual Grade School Math

“Nous évaluons les capacités de raisonnement de grands modèles linguistiques dans des contextes multilingues. Nous introduisons le benchmark Multilingual Grade School Math (MGSM), en traduisant manuellement 250 problèmes de mathématiques d'école primaire de l'ensemble de données GSM8K [...] dans dix langues typologiquement différentes.”

[\[MGSM 2022\]](#)

- **Format** : 250 problèmes de mathématiques traduits du GSM8K.
- **Niveaux** : Il se trouve autour de l'école primaire.
- **But** : Il s'agit de tester les capacités arithmétique sur 10 langues différentes qui sont le Bengali (BN), Chinese (ZH), French (FR), German (DE), Japanese (JA), Russian (RU), Spanish (ES), Swahili (SW), Telugu (TE), and Thai (TH).

Ceux qui créent des modèles de langage, fournissent la plupart du temps un document, introduisant le travail effectué et décrivant plus ou moins l'architecture du LLM. Nous disons “plus ou moins” car la notion de “open source” est un spectre potentiellement influencé par les enjeux de pouvoir et d'argent autour des modèles de langage puisque un nombre important de services incluant des LLM sont payants.

En général, ces personnes fournissent un benchmark. Des benchmarks qui deviennent petit à petit obsolètes et qui, par le fait qu'ils sont majoritairement open source, auraient pu être intégrés dans l'apprentissage initial du LLM, ce qui fausse les résultats. Nous ajouterons finalement que le multilinguisme des jeux de données est à modérer car plusieurs milliers de langues existent et les datasets ne frôlent pas ce nombre.

Malheureusement, faire des milliers de nouvelles tâches puis les vérifier et les normaliser en évitant tout biais est fastidieux. Il est imaginable qu'une entreprise privée ne veuille pas dépenser de l'argent pour des méthodes d'évaluation qui mèneront probablement à une remise en cause des capacités du produit qu'elles essaient de vendre alors que la plupart des benchmarks sont gratuits et de licence libre.

Une étude plus complète et critique faite par un organisme externe semble être aussi une méthode d'évaluation viable.

2.2 Une étude complète

L'étude complète telle qu'on se la représenterait couvrir un nombre suffisant de toutes les tâches possibles dans toutes les langues du monde (plus ou moins 7000) avec les métriques d'évaluations les plus pertinentes dans des formats variés. Cette étude doit aussi avoir correctement théorisé et défini les notions complexes que l'on mesure comme la logique, le raisonnement ou le bon sens.

Certaines études ont essayé de s'approcher de cet idéal exigeant pour apporter une évaluation plus approfondie.

Dans l'étude "*A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity*" [[MMMEC 2023](#)] ChatGPT été évalué d'une façon plus poussée que celle des benchmarks que nous avons vu.

Elle se démarque par les aspects suivants :

- **Multitâche** : L'étude a utilisé 21 jeux de données couvrant 8 tâches principales du TAL.
- **Multilingue** : L'étude teste l'analyse de sentiment, la traduction dans les deux sens avec l'anglais, la détection de langue. Le tout sur 8 langues regroupées selon leur présence en pourcentage dans le corpus d'apprentissage. Il y a les "HRL" langues à ressources élevées, les "MRL" langues à ressources moyennes et pour finir les "LRL" langue à ressources basses voire extrêmement basses (même vides) "XLR-L".

Language	#Speakers	CC Size (%)	Language Category
English (eng)	1.452B	46.320	HRL
Chinese (zho)	1.118B	4.837	HRL
French (fra)	235M	4.604	HRL
Indonesian (ind)	199M	0.781	MRL
Korean (kor)	81.7M	0.679	MRL
Javanese (jav)	68.3M	0.002	LRL
Sundanese (sun)	32.4M	0.001	LRL
Buginese (bug)	5M	0.000	X-LRL

Figure 3 - Tableau du multilinguisme du benchmark Multitask, Multilingual, Multimodal

- **Raisonnement théorisé** : Les 634 questions dites de raisonnement ont été regroupées selon le type de raisonnement. L'étude distingue le raisonnement déductif, inductif, abductif, mathématique, temporel, spatial, de sens commun, causal, analogique et multi-saut. Ils considèrent aussi différents niveaux de difficultés.
- **Test de l'amélioration des réponses via le prompting** : L'interaction au tour par tour entre le prompt et la réponse du modèle a été exploitée pour chercher à améliorer le résultat pour les résumés et la traduction.

D'autres études se servent d'humains pour juger les réponses de ChatGPT car certaines réponses ne sont pas facilement classables automatiquement comme bonnes ou mauvaises. [\[ACL 2023\]](#)

Ces études offrent un éventail de tâches évaluées qui sont certes non exhaustives mais qui ont le mérite d'explicitier la moindre démarche d'évaluation pour encourager la mise en place et l'amélioration de celles-ci. Des évaluations qui semblent de plus en plus essentielles pour guider voire réguler l'essor des LLM.

Ces deux études avaient trouvé des forces et des limites à ChatGPT avec des pourcentages de précisions. Elles avaient aussi décrit le potentiel de ChatGPT dans l'interaction au tour par tour. Une interaction proche des conditions dites réelles de l'humain qui "prompt" (requête) au programme informatique plus d'une fois si la réponse ne le satisfait pas cherchant le meilleur résultat possible.

Le benchmark **MINT** présenté en mars 2024 exploite ce potentiel en fournissant une évaluation avec de multiples datasets (code, prise de décision et raisonnement dans un datasets compressé de 586 tâches) comme les précédents benchmark mais en donnant jusqu'à 5 chances aux LLM évalués pour améliorer la réponse. [MINT 2023]

Les modèles avaient l'occasion d'améliorer les performances en leur renvoyant les requêtes et réponses avec des commentaires et critiques (**feed-back**).

D'après leurs résultats il y a bien un potentiel d'amélioration pour les LLM .

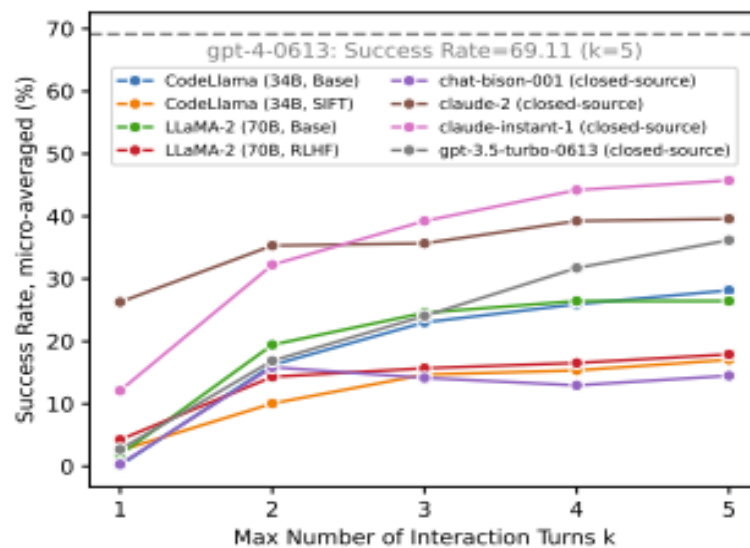


Figure 4 - Graphique représentant l'évolution des performances selon le nombre d'interaction

Cependant dû aux moyens de cette étude, les critiques et commentaires en prompt étaient faits par GPT-4 et non des humains. L'interaction est celle d'un modèle avec un autre modèle plus performant. Cela permet de fournir une quantité de résultats importante mais ne reflète pas forcément le potentiel en utilisation concrète par l'humain.

De plus, nous pouvons être amenés à nous demander si tous les humains font leur requête comme les énoncés des tâches présentes dans les jeux de données savamment sélectionnés/écrits/traduits/raffinés par des chercheurs ayant un bon niveau de la langue standard. La réponse est probablement non au vu de la variabilité linguistique de chaque individu.

Des conditions réelles d'interactions lors de l'évaluation des LLM pourraient indiquer leurs capacités dans un usage naturel mais aussi dégager des tendances et biais dans l'utilisation par les humains.

2.3 Une étude centrée sur les humains

Dans l'étude de Harvard Business School, il a été demandé aux **758** consultants d'un même cabinet à Boston de réaliser 18 tâches d'expertise en temps limité qu'ils sont souvent amenés à faire dans leur quotidien professionnel. [[NJTF 2023](#)]

Trois groupe ont été formés :

1. Un groupe sans aide particulière.
2. Un groupe avec l'accès à ChatGPT-4 (non obligatoire).
3. Un groupe avec l'accès à ChatGPT-4 (non obligatoire) et des outils pour mieux se servir de l'IA appelés "prompt engineering". (vidéos et documents pédagogiques pour des stratégies d'usages).

Les tâches ont été sélectionnées de sorte qu'il y ait une partie que l'IA réussit à traiter dite **"dans la frontière"** et une partie qu'elle peine à résoudre via un copié-collé du problème dite **"hors de la frontière"**. Elles sont notées par des humains et ChatGPT-4.

Les chercheurs/ses ont finalement distingué des profils comportementaux face à une tâche lors de l'accès à ChatGPT :

- **Centaure** : Ce sont les consultants qui ont utilisé l'outil uniquement pour certaines sous-tâches individuelles et non la tâche en entier. Ils divisent le travail entre eux et le modèle.
- **Cyborg** : Ce sont les consultants qui ont continuellement utilisé l'outil sur la tâche complète en interaction pour améliorer les résultats. Ils ont intégré le modèle dans tout leur flux de travail.

Ces deux profils ne sont pas mesurées ni quantifiés concrètement mais d'autres observations faites par l'étude semblent plus explicites :

1. Pour les tâches hors de la frontière, les deux groupes de consultants qui avaient accès à ChatGPT ont eu de moins bons résultats.
2. Pour les tâches dans la frontière, l'utilisation de ChatGPT donne des résultats supérieurs.
3. Le groupe avec l'outil et les éléments pour aider au prompting ont eu de meilleurs résultats que celui qui avait juste ChatGPT.
4. Les participants ayant le moins utilisé contenu des réponses de l'outil dans leurs réponses ont eu en moyenne de meilleurs résultats.

Il faut cependant préciser qu'entre les deux frontières, les métriques d'évaluation et les types de questions ne sont pas les mêmes. De plus, pour les tâches à l'intérieur de la frontière la qualité de la réponse est notée de 1 à 5 par des humains et ChatGPT-4.

Quelles compétences ont-ils pour pouvoir noter objectivement et impartialement ?

Cela démontre toute la difficulté dans l'évaluation de textes générés qui relève parfois de la subjectivité. Qu'est-ce qu'un bon résumé ?

Pour synthétiser, toutes ces études proposent des méthodes d'évaluation intéressantes que ce soit dans leurs métriques d'évaluations (le score BLEU par exemple) [[BLEU 2002](#)] , la diversité des tâches ou le protocole mis en place (faire plusieurs groupes avec un groupe témoin).

Nous n'avons cependant pas trouvé d'études analysant de manière plus poussée les prompts des humains en conditions réelles sur des tâches différentes. Nous pensons qu'évaluer conjointement l'humain et la machine en contexte réel pourrait apporter une évaluation qualitative.

Il est intéressant de se demander si l'humain se rend compte ou a conscience des capacités et limites de son outil pour une utilisation pertinente.

C'est en nous inspirant de ces études que nous avons voulu apporter notre pierre à l'édifice en menant notre propre expérience avec les moyens dont nous disposions. Cette expérience a abouti à l'élaboration d'un corpus.

3. LE CORPUS

Nous présentons maintenant notre corpus élaboré suite à une expérience menée sur **62** étudiants de L1 Science du Langage. Ceux-ci ont passé un premier test seuls puis le même test une deuxième fois avec la possibilité d'utiliser ChatGPT.

Cette partie définit l'objectif du corpus, son contenu, la description des tâches qui le composent et finalement la méthodologie qui a été employée pour exploiter les données brutes.

3.1 But du corpus

Avant d'entamer la description de notre corpus, il convient de recentrer le propos sur des hypothèses qui pourraient être validées ou réfutées par l'étude du corpus.

1. **ChatGPT améliore les résultats en conditions réelles d'interaction avec l'humain.**
2. **Les étudiants ont des stratégies de prompting qui influencent la performance.**
3. **Les étudiants ont conscience des capacités et limites de ChatGPT.**
4. **Les étudiants se reposent trop sur l'outil.**

3.2 Contenu du corpus

Le corpus est organisé dans des tableurs (fichiers .xlsx) et documents de textes (fichiers .txt) avec les données suivantes :

- **20 énoncés** de tâches donnant au total **23 questions** à répondre (voir [annexe 3](#))
- **764 Requêtes** écrites par des étudiants destinées à ChatGPT avec le numéro de la tâche sur laquelle porte la requête. (exemple à l'[annexe 4](#))
- **764 Réponses** à ces requêtes par ChatGPT avec le numéro de la tâche sur laquelle porte la réponse. (exemple à l'[annexe 4](#))
- **1467 réponses** écrites par **62 étudiants** sur une copie en papier (puis saisies sur tableur) répondant aux 23 questions.
- **1535 réponses** issues d'un formulaire numérique rempli par les mêmes 62 étudiants répondant aux 23 questions.

3.3 Description des tâches du corpus

Nous avons établi **un ensemble de 20 tâches** de natures différentes en français en essayant de faire varier la difficulté des questions selon notre propre opinion et celui de proches à qui nous avons préalablement soumis le test.

Pour les compréhensions écrites il y avait deux questions par tâches d'où le **nombre de questions** de **23** au lieu de 20.

Une tâche correspond à un énoncé pouvant donner lieu à une ou plusieurs questions à répondre. L'ensemble des énoncés se trouve à l'[annexe 3](#).

Distribution typologique des tâches :

- 3 compréhensions écrites (QCM) : 2 questions par tâche.
- 2 expressions écrites (réponse à rédiger) : 1 question par tâche.
- 4 problèmes mathématiques (réponse à rédiger) : 1 question par tâche.
- 4 problèmes de logique (2 QCM + 2 réponses à rédiger) : 1 question par tâche.
- 7 questions de connaissance (QCM) : 1 question par tâche.

3.4 Méthodologie d'exploitation du corpus

Nous décrivons ici les premiers traitements qui nous ont permis d'exploiter les données du corpus. Les traitements sont différents selon le type de donnée.

1. Traitement des requêtes et réponses récupérées sur ChatGPT

- a. **Segmentation** du contenu du fichier des requêtes/réponses grâce à des balises insérées pendant la saisie manuelle de données qui permettent de savoir sur quelles questions les requêtes et réponses portent.
- b. **Nettoyage** de caractères indésirables.
- c. **Tokenisation** avec la librairie *spacy* sur python.

2. Étiquetage des réponses (selon leur format) aux questions. Elles sont notées **correctes ou fausses** via un algorithme sur **python** (**hors tâches d'expressions écrites**)

- a. **QCM** : Il n'y a qu'un seul choix qui est correct. (Le nombre de choix varie de 4 à 5)
- b. **Réponse à rédiger** : Il n'y a qu'une seule réponse possible avec une tolérance à certaines variations orthographiques ou de notations mathématiques.

Tous les étudiants ont ensuite été **anonymisés**.

3.5 Les métriques d'évaluation

Une métrique ici peut être une formule, une technique ou un calcul dont le but est de quantifier la performance. Nous listons les métriques utilisées pour différentes données.

Le taux de réussite d'un étudiant à un test :

Il s'agit de la proportion de bonnes réponses sur le nombre de questions totales pouvant être considérées vraies ou fausses (**c'est -à -dire hors expressions écrites**). Nous avons choisi de parler en pourcentage pour des visualisations plus faciles.

Un étudiant ayant 9 bonnes réponses aura donc un taux de réussite d'environ :
 $(9/21)*100 = 42,85\%$

La similarité cosinus :

La similarité cosinus est une technique qui permet de mesurer à quel point deux textes sont similaires. Plus la valeur est proche de 1, plus les textes se ressemblent [[SIM COS 2013](#)]. Nous pouvons l'appliquer aux requêtes, réponses de ChatGPT et expressions écrites.

Le score BLEU :

Le score BLEU mesure à quel point un texte généré est similaire à un autre texte en comparant les séquences de mots (n-grammes) entre eux. En utilisant la librairie Python *nltk*, on peut calculer ce score pour évaluer la similarité entre les deux textes. Plus le score est élevé, plus les textes sont proches. Il est souvent utilisé pour l'évaluation de traductions, résumés ou autres textes générés. [[BLEU 2002](#)]

Nous allons maintenant passer à la présentation de l'expérience qui a permis l'élaboration de ce corpus.

4. L'EXPÉRIENCE

Nous décrivons ici le contexte de l'expérience, le choix de l'outil et le déroulement de l'expérience.

4.1 Contexte de l'expérience

Nous avons eu l'opportunité de mener une expérience le **26 avril 2024** avec des étudiants de première année de la licence Sciences du Langage à la Sorbonne, en simulant des conditions d'examen.

Par la nature des conditions de l'expérience nous avons des contraintes :

- Nombre d'étudiants limité à la promotion.
- Temps limité d'accès aux salles où nous avons fait l'expérience.
- Pas d'argent pour avoir des serveurs ou du matériel facilitant l'utilisation d'un LLM.

4.2 Le choix de l'outil

Nous avons utilisé ChatGPT car il était gratuit et compatible avec nos moyens. En effet sa plateforme de chat permet de se reconnecter à tous les comptes des étudiants afin de récupérer les requêtes et réponses. Si nous avions eu des serveurs, nous aurions utilisé des modèles open sources beaucoup plus légers au niveau du nombre de paramètres.

On voit sur la *figure 5* que les pourcentages de réussite sont différents selon les LLM pour les 20 tâches (test répété 10 fois). ChatGPT a les meilleurs résultats avec un taux moyen de réussite de **68,75%**, cela s'explique en partie par son nombre de paramètres.

Nous avons évalué à la main toutes les réponses.

Les expressions écrites sont considérées comme correctes si elles ne sont pas hors sujet.

Pourcentage de réponses correctes selon les types de tâches

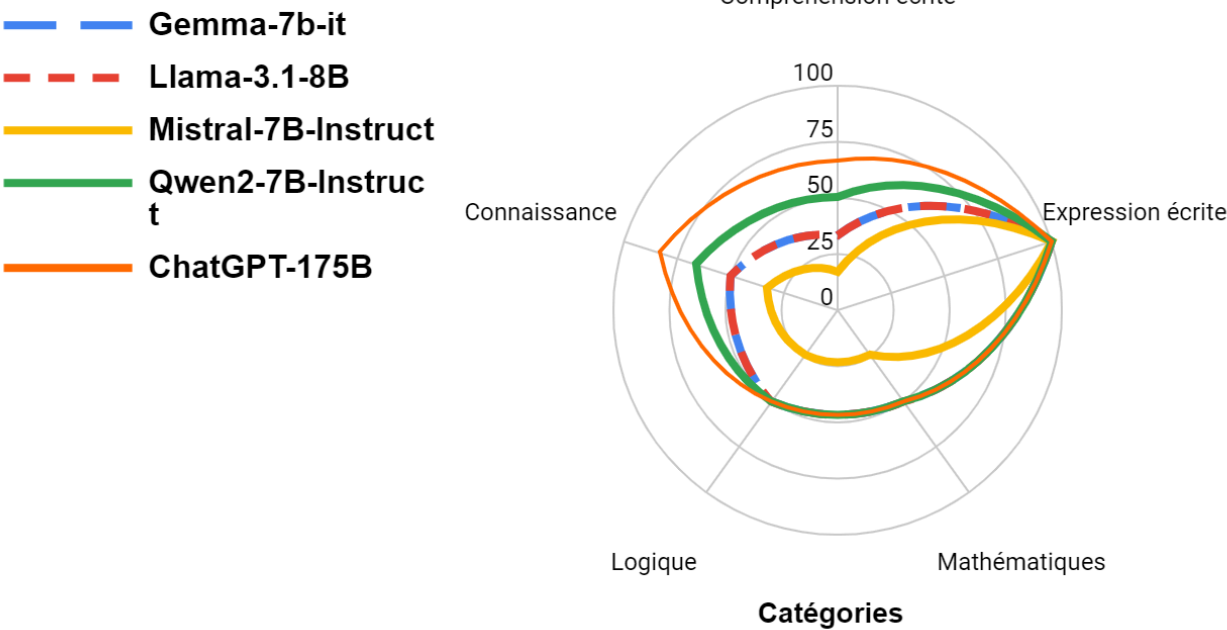


Figure 5 - Pourcentage de réponses correctes selon les types de tâches

	ChatGPT 175B	Gemma 7B-it	LLama 3.1-8B	Mistral 7B-Instruct	Qwen2 7B-Instruct
Compréhension écrite	66,66%	33,33%	33,33%	16,66%	50%
Expression écrite	100%	100%	100%	100%	100%
Mathématiques	50%	50%	50%	25%	75%
Logique	50%	50%	50%	25%	50%
Connaissance	83,33%	50%	50%	33,33%	66,66%

Figure 6 - Tableau récapitulatif de la figure 5

Présentons les outils que nous aurions pu choisir :

- **Gemma-7B-it** est un modèle créé par Google Deepmind [[GEMMA 2024](#)]. Ce modèle a une taille au total de 7 milliards de paramètres et est construit et entraîné principalement sur des données en anglais. Il a une taille de vocabulaire de 256000 et constitue un modèle entraîné sur 6To de token d'apprentissage. Il est principalement utilisé pour traiter des données en anglais. [[GT 2024](#)]
- **Llama-3.1-8B** est un modèle fondé par l'entreprise Meta en avril 2024 , ce modèle a une taille de vocabulaire de 128000 et a été entraîné sur 15 To de token d'apprentissage. Lorsque l'on pose la question au modèle, il affirme traiter des données de plusieurs langues différentes telles que l'Anglais , l'Allemand , Français , l'Italien , le Portugais , l'Hindi, l'Espagnol et le Thaïlandais. [[LAMA 2024](#)]
- **Mistral-7B-Instruct** est un modèle de LLM fondé par l'entreprise Mistral AI et a un nombre de paramètres de 7B avec une taille de vocabulaire de 32000. Ce modèle quant à lui affirme traiter un total de données de 5 langues différentes qui sont l'Anglais , le Français , l'Espagnol, l'Allemand et l'Italien. [[MISTRAL 2024](#)]
- **Qwen2-7B-Instruct** est un modèle créé par l'entreprise Alibaba Cloud avec 7 milliards de paramètres. [[QWEN 2024](#)]

Abordons maintenant le déroulement de l'expérience.

4.3 Déroulement de l'épreuve

Notre expérience s'est déroulée en deux tests identiques que tous les étudiants ont passés. Ces deux tests sont formés des 20 tâches du corpus.

1ère partie : Le test contrôle

Le test s'est déroulé dans l'amphithéâtre 128 à Malesherbes (un site de la Sorbonne) où les étudiants devaient composer sans l'aide d'aucune aide extérieure et dont la durée du test était de 45 minutes.

Un des 3 sujets différents leurs ont été distribués avec les mêmes questions (A , B et C) mais dans des ordres différents pour réduire la triche :

- Sujet A : L'ordre canonique des questions.
- Sujet B : L'ordre des questions a été répartie dans un ordre aléatoire.
- Sujet C : L'ordre des questions est inverse à celui du sujet A.

2e partie: Le test sur ordinateur avec ChatGPT

Le test s'est situé cette fois-ci en salle informatique (BS17-BS18 et salle 220) dans laquelle chaque étudiant avait un identifiant et un mot de passe unique pour se connecter au site web ChatGPT via les ordinateurs à disposition.

Ils avaient la possibilité cette fois-ci d'utiliser ou non l'outil.

Le contenu des questions qui a été donné était le même que celui du premier test sauf que cette fois-ci il s'agissait d'un format numérique (Google Forms).

5. RÉSULTATS

Nous montrons dans cette partie les résultats. Ils nous donnent des informations du point de la réussite, des questions, des étudiants et de l'interaction quantitative et qualitative.

Nous rappelons que les résultats concernent **62 étudiants** ayant tous passé le test sans aide puis le test avec ChatGPT à disposition. Ils ont des identifiants chiffrés pour être anonymes.

Comme mentionné dans la description du corpus, les pourcentages de réussites ne prennent pas en compte les deux tâches d'expression écrite. Nous sommes donc en train d'évaluer **18 tâches** donnant lieu à **21 questions**.

Le premier test sans aide s'appelle "**Test humain**" et le deuxième avec ChatGPT s'appelle "**Test humain + ChatGPT**".

5.1 La réussite : ChatGPT améliore-t-il les résultats ?

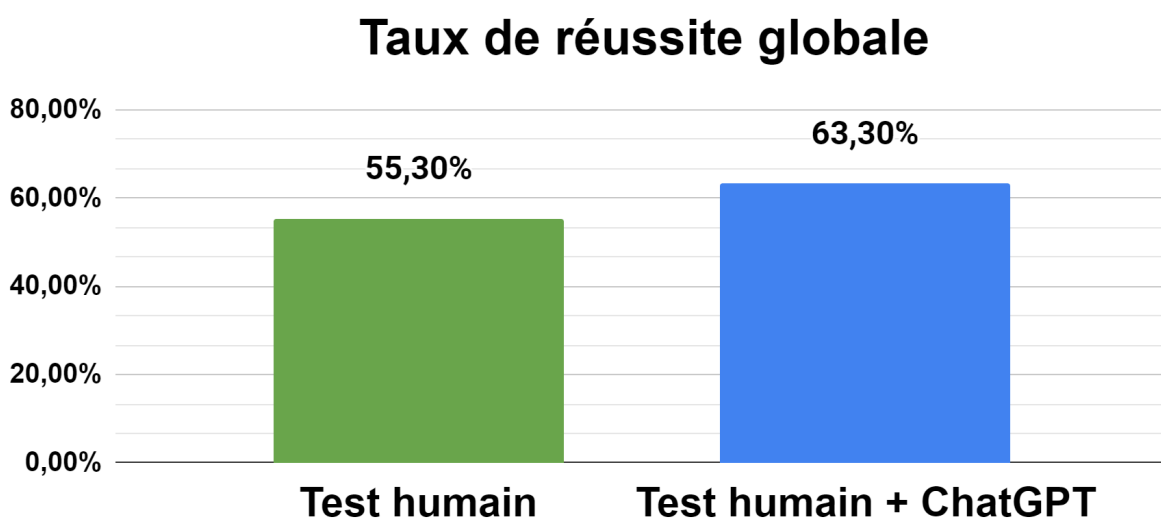


Figure 7 - Taux de réussite globale

Observation de la figure 7 :

Nous pouvons voir qu'au test humain + ChatGPT (test sur machine), il y a une amélioration globale des réponses correctes de 8% par rapport au test humain. Pour un total de 21 questions 8% représenterait une question et demi réussite en plus. Ce qui est étonnant c'est que ChatGPT seul atteint un taux de réussite globale de 68,75% ce qui est supérieur aux

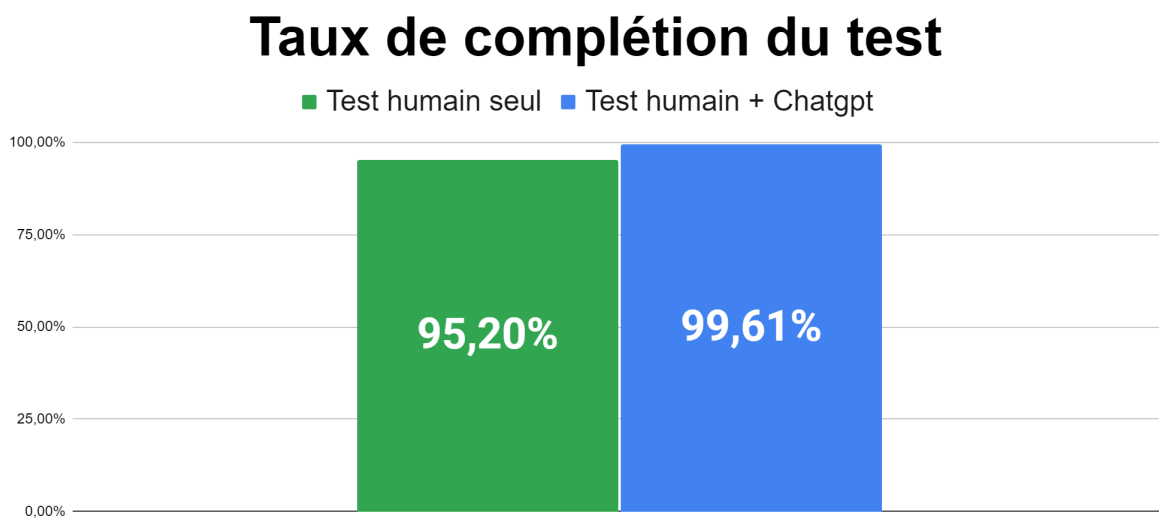


Figure 8 - Taux de complétion du test

Observation de la figure 8 :

On voit sur la *figure 8* que les élèves ont en moyenne répondu à plus de questions au test humain + ChatGPT qu'au test humain. Cela peut être lié au boost de productivité qu'apporterait ChatGPT. Nous pensons plutôt que c'est parce que les étudiants ont pu aller plus vite sur ce deuxième test puisqu'ils venaient d'être confrontés aux mêmes questions au test humain.

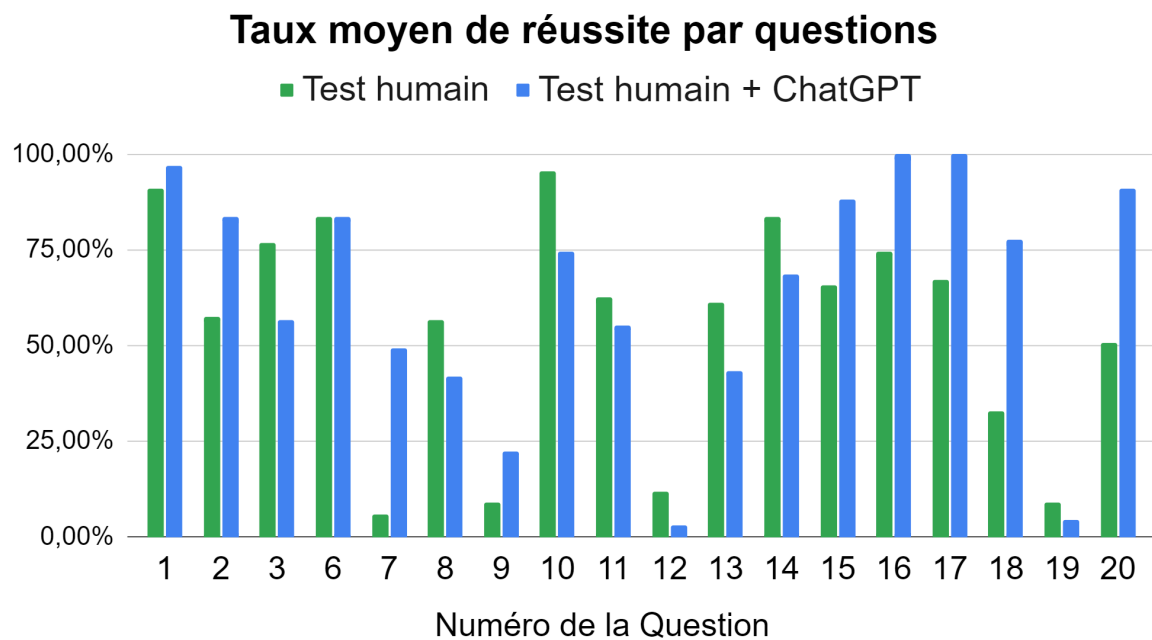


Figure 9 - Graphique comparatif des taux de réussite par question des tests

Observation de la figure 9 :

Sur la *figure 9* nous observons que malgré l'écart de **8%** du taux de réussite global entre le test humain et le test humain + ChatGPT, il y a tout de même plusieurs questions qui ont été mieux réussies lors du test humain.

De plus, si nous comptons, il y a autant de questions mieux réussies dans le test humain que dans le test humain + ChatGPT. La présence de ChatGPT comme outil disponible n'a donc pas été bénéfique sur l'intégralité des questions.

Interprétation générale sur la réussite :

En prenant les taux de réussite globaux des deux tests, nous avons une moyenne d'environ **59,3%**. De ce fait, on ne peut pas dire que ce test était facile pour les étudiants.

Ces résultats tendent à confirmer **l'hypothèse 1** affirmant que l'étudiant a de meilleurs résultats avec ChatGPT à disposition. Mais nous voyons bien que cela ne s'applique pas à toutes les tâches.

De par la disparité entre les des taux de réussite selon les questions, il est intéressant de regarder les résultats du point de vue des questions et des types de questions.

5.2 Les questions

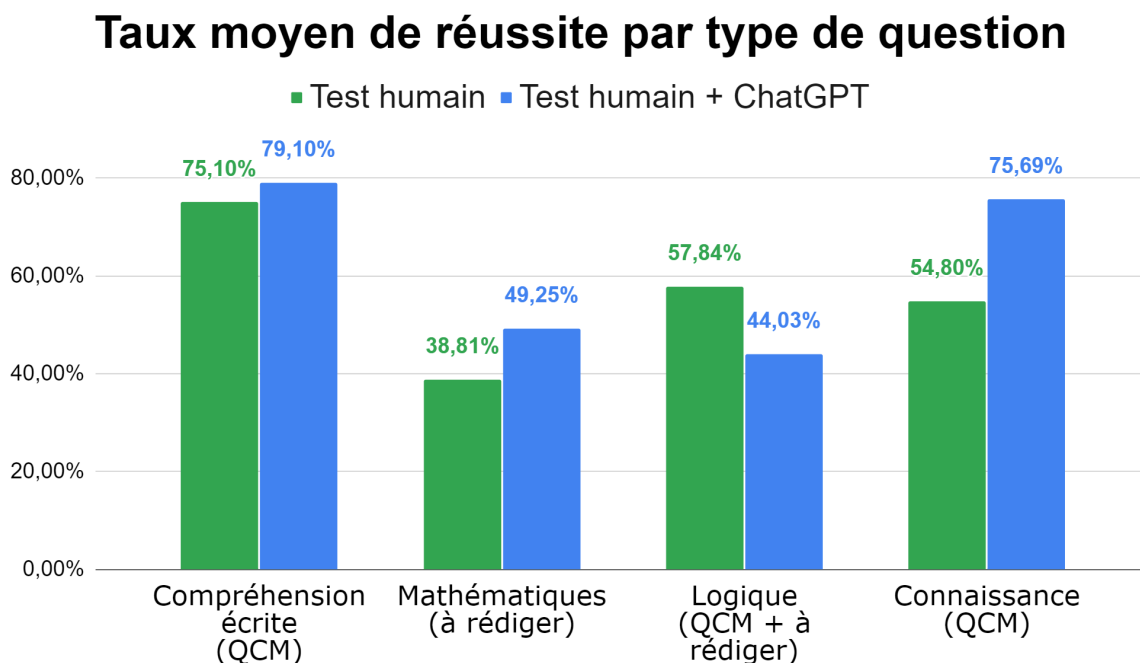


Figure 10 - Taux moyen de réussite par type de question

Observation de la figure 10 :

Nous constatons que la présence de ChatGPT améliore notablement les performances pour les tâches de connaissance et dégrade les performances pour les tâches de logique.

Nous pouvons aussi voir que les deux catégories avec les taux les plus bas au test humain + ChatGPT comprennent des questions à rédiger. Le format des réponses pourrait aussi avoir un lien avec la réussite.

Le cas de la question 8 (mathématiques)

Prenons la question 8 comme exemple.

Question	Test humain	Humain + ChatGPT
8	56,72%	41,79%

Figure 11 - Tableau récapitulatif du taux de réussite de la question 8

Énoncé de la question 8 :

“ Avant-hier je n'avais pas de pommes, aujourd'hui j'en ai gagné 16, hier j'ai gagné le huitième de ce que j'ai gagné aujourd'hui et demain j'en gagnerai le triple. 1- Combien ai-je de pomme au total pour aujourd'hui ? (donnez juste le nombre/chiffre): ”

Nous considérons cette question comme facile pour des étudiants universitaires.

ChatGPT ne réussit pas cette question lorsque nous copions l'énoncé en prompt.

Cela peut expliquer le fait que le taux de réussite au test humain + ChatGPT soit plus bas. Cela voudrait dire que certains étudiants ont plus fait confiance à ChatGPT pour cette tâche de mathématique ce qui les aurait amené à échouer à une question qu'ils avaient réussi au test d'avant (test humain).

Distribution du nombre de requêtes par type de tâches

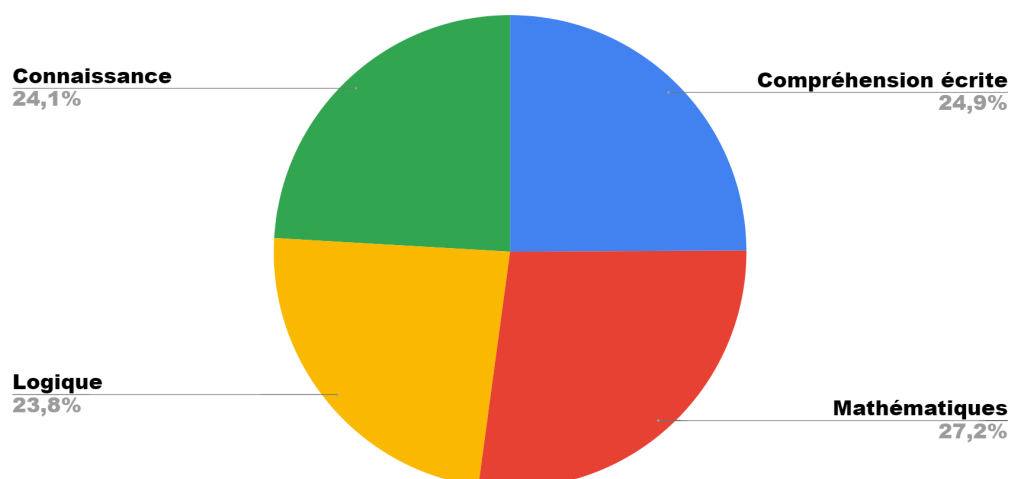


Figure 12 - Distribution du nombre de requêtes par type de tâches

Observation de la figure 12 :

On voit que le nombre de requêtes envoyées à ChatGPT lors du test humain + ChatGPT se répartit de façon équivalente parmi les différents types de tâches. Pourtant ChatGPT ne réussit pas bien les différentes tâches également comme on le voit *figure 13*.

Type de tâche	ChatGPT seul (10 itérations)	Test humain + ChatGPT
Compréhension écrite	66,66%	79,10%
Mathématiques	50%	49,25%
Logiques	50%	44,03%
Connaissance	83,33%	75,69%

Figure 13 - Tableau comparatif du taux de réussite selon le type de tâche ChatGPT seul / test humain + ChatGPT

Interprétation générale sur les questions :

ChatGPT semble améliorer les résultats au test humain + ChatGPT dans la limite de ses propres capacités montrées *figure 13*. Les étudiants ont régressé au test humain + ChatGPT en logique, logique peu réussi par ChatGPT seul.

Ces résultats participent à réfuter l'hypothèse 4 “**les étudiants ont conscience des capacités et limites de ChatGPT**” car ils ont en général autant fait de requêtes pour des questions que ChatGPT réussit bien seul que pour des questions qu’il réussit bien.

5.3 Les étudiants

taux de réussite le plus bas d'un élève :

Humain seul : 28,6%

Humain avec ChatGPT : **38,1%**

taux de réussite le plus élevé d'un élève :

Humain seul : 85,7%

Humain avec ChatGPT : **92,24%**

Comparaison du taux de réussite par étudiants

étudiants classés dans l'ordre décroissant du taux de réussite au test humain seul

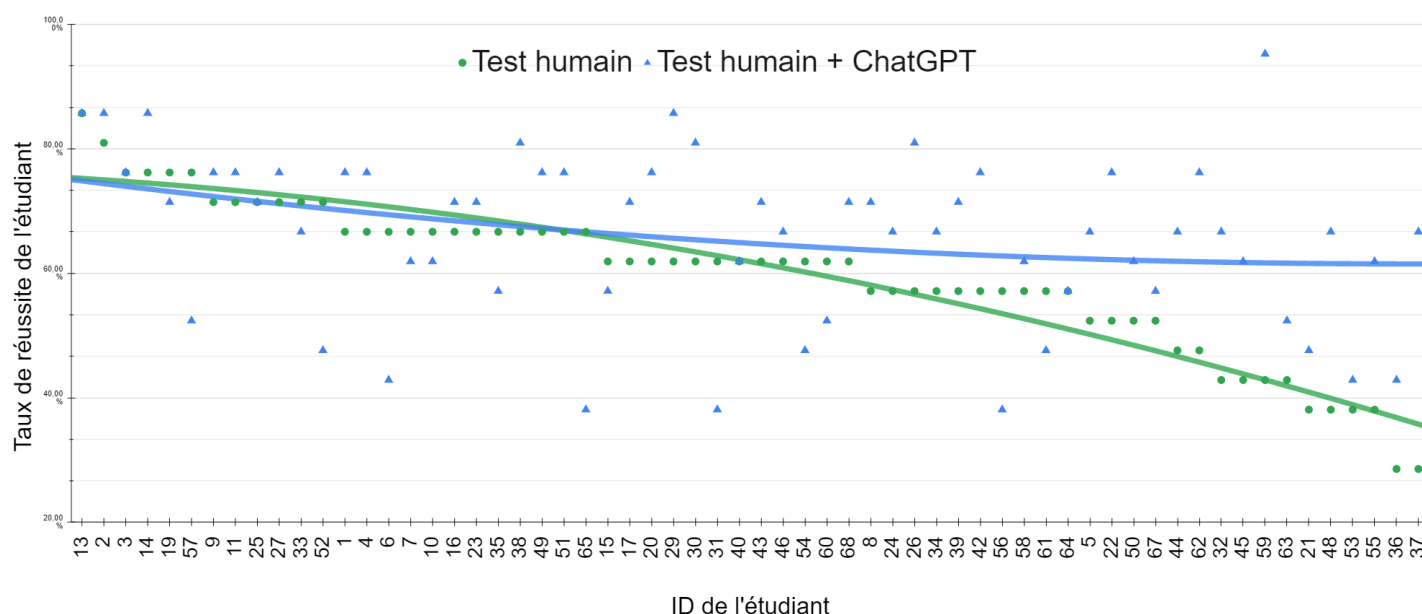


Figure 14 - Comparaison du taux de réussite par étudiants

Observation de la figure 14:

Ce graphique montre les deux taux de réussite par étudiant.

Les deux courbes de tendance nous montrent qu'en général plus un étudiant a eu des résultats bas au test humain, plus il va s'améliorer au test humain + ChatGPT.

Néanmoins, le taux de réussite au test humain + ChatGPT ne subit pas une augmentation au fur et à mesure que l'on décroît dans le taux de réussite au test humain.

Il faut également dire qu'il y a beaucoup de points éloignés de la courbe de tendance pour le test humain + ChatGPT ce qui la rend moins représentative de la réalité.

On remarque qu'environ **67% des étudiants progressent** du test humain au test humain + ChatGPT contre **24% qui régressent** (avec 9% des élèves qui stagne).

Interprétation générale sur les étudiants:

Le dernier quart des étudiants ayant les moins bons résultats au test humain ont tous progressé au test humain + ChatGPT. Pour ce qui est du reste, il y a eu des progressions et régressions d'ampleurs diverses non corrélées à la note au test humain.

Nous pouvons distinguer les deux profils d'étudiant :

- Un étudiant qui s'est amélioré avec ChatGPT.
- Un étudiant qui a régressé avec ChatGPT.

Voyons si les requêtes ont une influence sur l'évolution de la réussite d'un test à l'autre, car la différence de résultat est peut-être liée au fait qu'ils passent deux fois le même test. Rien ne dit pour l'instant qu'ils ont vraiment utilisé l'outil puisqu'ils n'étaient pas obligés de le faire.

5.4 L'interaction quantitative

- **45** étudiants sur les **62** ont choisi d'utiliser ChatGPT lors du test humain + ChatGPT soit **72%** de taux d'utilisation de ChatGPT or il y a 9% des étudiants qui ont eu des résultats identiques aux deux test, cela veut dire que certains étudiants n'ont pas mis les mêmes réponses aux questions sans utiliser ChatGPT.

Total des requêtes	Maximum de requêtes par étudiant	Minimum de requêtes par étudiant	Moyenne de requêtes par étudiant	Médiane de requêtes par étudiant
764	79	0	12	11

Figure 15 - Tableau des statistiques générales sur les requêtes

Observation de la figure 15 :

Le fait le plus marquant est la différence entre la moyenne de requêtes par étudiant de **12** et le maximum de **79**. Cela dénote des comportements hétérogènes des étudiants face à aux machines.

Regardons si nous pouvons faire un lien avec la progression de la réussite depuis le test humain vers le test humain + ChatGPT.

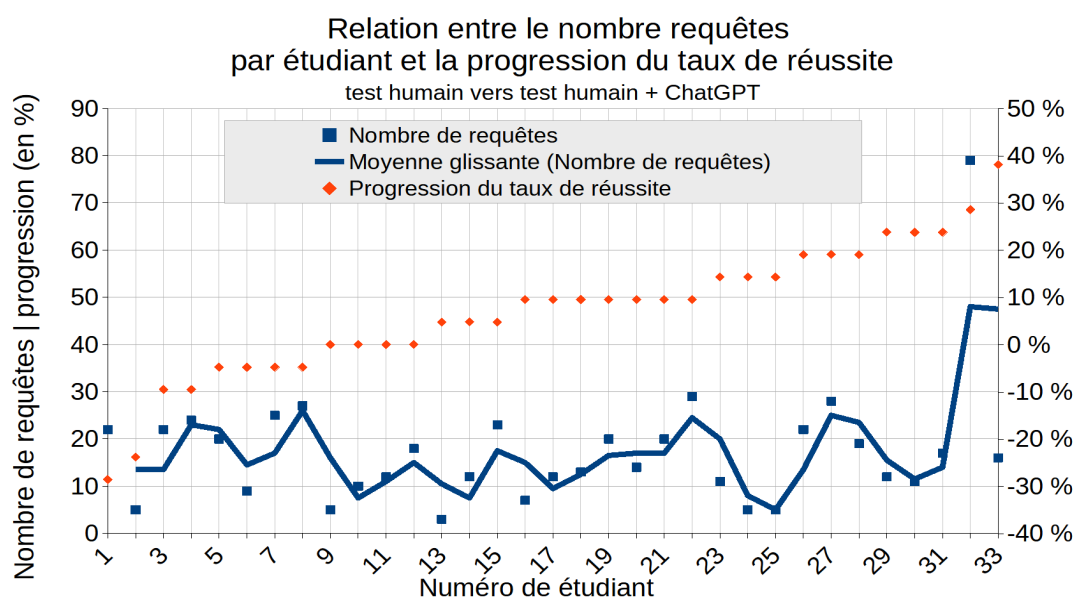


Figure 16 - Relation entre le nombre de requêtes par étudiant et la progression du taux de réussite

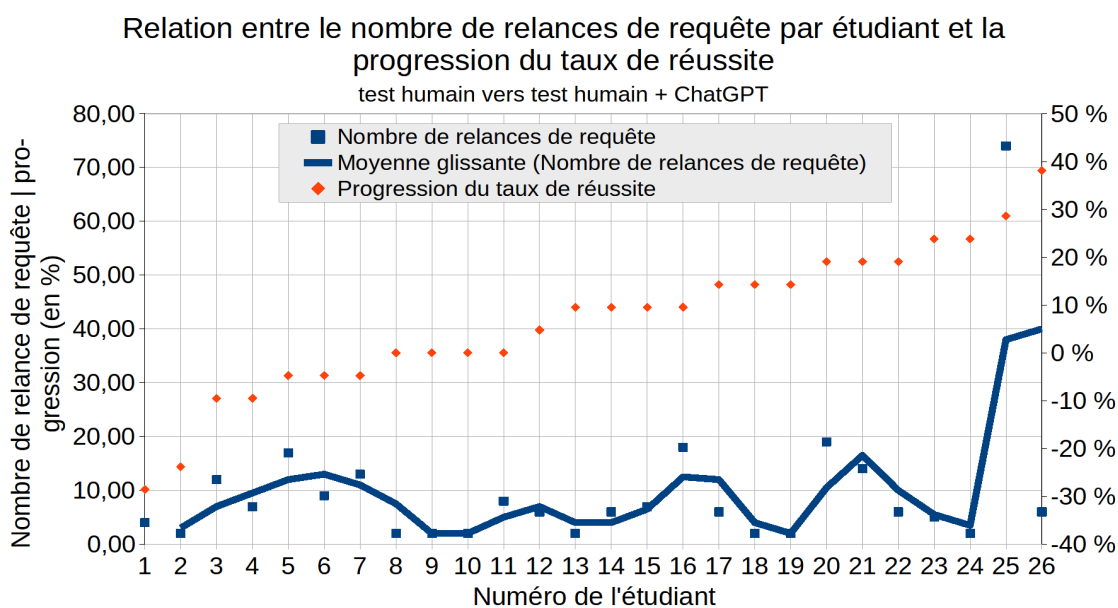


Figure 17 - Relation entre le nombre de relances de requête par étudiant et la progression du taux de réussite

Observation des figures 16 et 17 :

Les courbes de tendance pour le nombre de requêtes ou le nombre de relances de requête (une relance est une ou plusieurs requêtes qui suivent la première pour une question) ne

montrent pas de lien entre la quantité de requête et la progression du taux de réussite depuis le test humain vers le test humain + ChatGPT.

Pourtant, nous constatons tout de même sur la *figure 17* que pour le groupe des 45 étudiants qui ont utilisé ChatGPT, les élèves ayant régressé ont eu tendance à plus interagir avec ChatGPT comparés au groupe ayant progressé.

Nombre moyen de requêtes des élèves ayant régressé	Nombre médian de requêtes des élèves ayant régressé	Nombre moyen de requêtes des élèves ayant progressé	Nombre médian de requêtes des élèves ayant progressé
17	20	10	5

Figure 18 - Tableau statistique comparatif sur le nombre de requêtes

Interprétation générale sur les interactions quantitatives :

Les étudiants n'ont pas interagi dans les mêmes proportions et ces proportions ne semblent ne pas avoir d'effet sur la progression du taux de réussite du test humain vers le test humain + ChatGPT.

Les résultats mènent à nuancer l'hypothèse 5 “ **les étudiants se reposent trop sur l'outil** ” car un étudiant prompt 12 fois en moyenne sur 23 questions.

Les élèves ont effectivement des comportements d'utilisation de l'outil différents d'un point de vue quantitatif ce qui semblerait valider l'hypothèse 3 “ **Il y a des profils comportementaux avec l'outil qui se distinguent** ”.

Néanmoins on ne peut pas vraiment parler de “profils” simplement en constatant des différences surtout si elles ne peuvent être regroupées.

Regardons si ces interactions sont également différentes.

5.5 L'interaction qualitative

Relation entre la progression du taux de réussite et l'originalité de la requête par étudiant

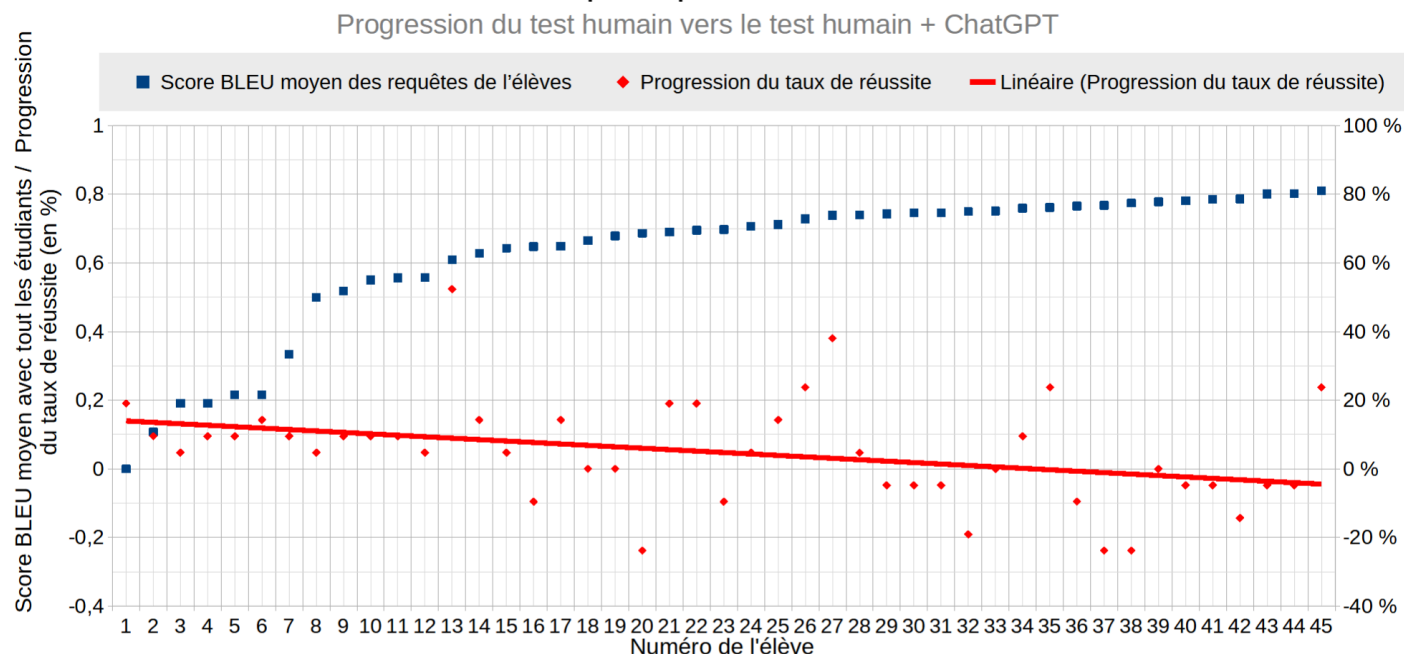


Figure 19 - Relation entre la progression du taux de réussite et l'originalité de la requête par étudiant

Observation de la figure 19 :

Nous avons calculé pour chaque étudiant la moyenne des scores BLEU établie à chaque questions entre les requêtes de tous les autres étudiants.

Plus un étudiant a un score BLEU proche de 1, plus ses requêtes ont été similaires à celles des autres étudiants.

La courbe nous montre que les étudiants les plus similaires aux autres dans leurs requêtes ont tendance à avoir une progression du test humain vers le test humain + ChatGPT moins élevé. Cependant il y a plusieurs points éloignés de cette courbe de tendance ce qui limite son aspect représentatif.

Si nous considérons qu'un étudiant est assez similaire aux autres dans ses requêtes à partir d'un score BLEU supérieur ou égal à 0,75, alors **13 étudiants sont considérés comme similaires aux autres dans leurs requêtes.**

Ces étudiants ont une moyenne de progressions de taux de réussite du test humain au test humain + ChatGPT de **-2,6% contre 7,74%** Pour les étudiant avec un score BLEU moyen inférieur à 0,75.

Regardons si les élèves assez sont similaires dans leurs prompts parce qu'ils ont copié l'énoncé.

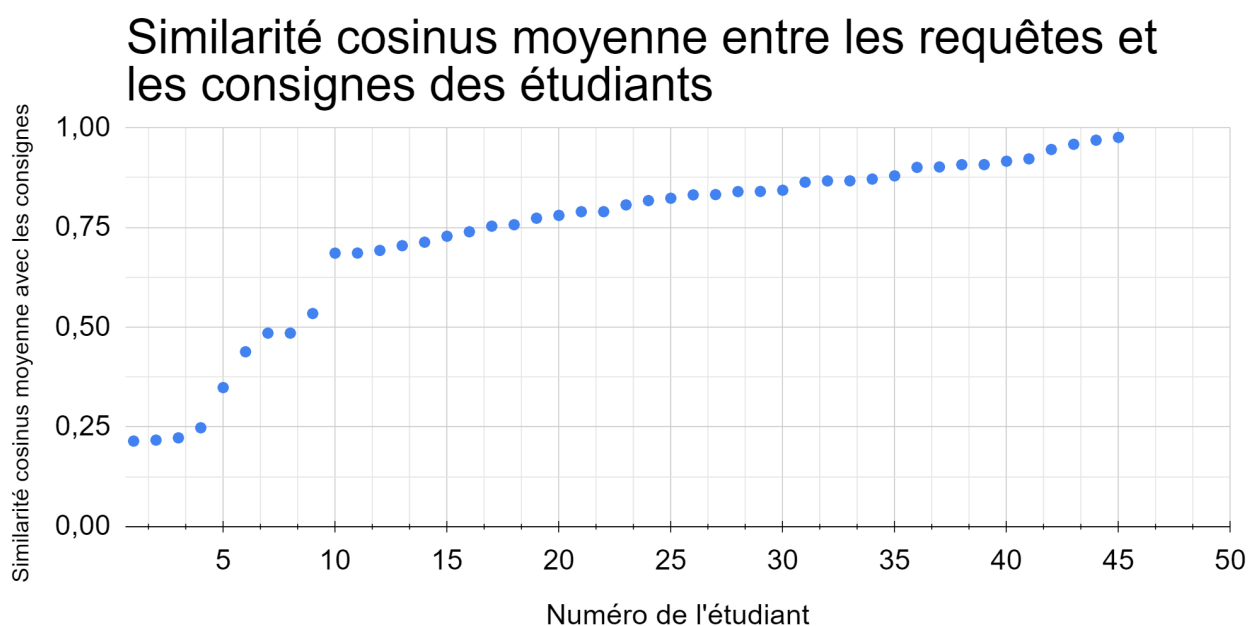


Figure 20 - Similarité cosinus moyenne entre les requêtes et les consignes des étudiants

Observation de la figure 20 :

Nous constatons que plusieurs étudiants ont eu tendance à copier la consigne tandis que d'autres ont préféré prendre une petite partie de la consigne ou au contraire écrire eux même pour faire leur requêtes.

Interprétation générale sur les interactions qualitatives :

Les étudiants ont en majorité préféré rester assez fidèles à l'énoncé dans leur stratégie de prompting. Les énoncés étant assez correctement rédigés, il est cohérent de les avoir copiés.

5.6 Discussion et limites

Les résultats nous ont montré qu'en présence de ChatGPT, les étudiants ont bénéficié en moyenne d'une légère amélioration de leurs résultats, mais essentiellement pour des tâches de connaissance. Les étudiants ne semblent pas avoir discerné les forces et limites de ChatGPT en ayant fait des requêtes à parts égales dans les différents types de tâche alors même que certaines étaient peu réussies par ChatGPT seul. Certaines étaient pourtant selon notre opinion assez faciles pour un niveau universitaire comme la question 8.

Ces résultats nous ont aussi montré que les étudiants ont des comportements différents avec l'outil ne l'utilisant pas du tout dans certains cas. Des comportements qui se reflètent ailleurs comme dans le nombre de requête ou les stratégie de prompting (copier l'énoncé ou écrire soit même son prompt) et que cela peut influencer les résultats.

Cependant, les conditions d'expériences ont limité le nombre d'étudiants participant aux tests et le nombre de questions des deux tests. Nous n'avons pas pu distinguer clairement des catégories dans les types de comportements devant ChatGPT.

Évaluer sur un nombre insuffisant de questions comparé aux centaines voir milliers présentes dans les benchmarks classiques réduit la portée de cette étude.

Nous avons exploré les données du corpus avec un certain point de vue, il n'est donc pas impossible que d'autres façons d'exploiter les données donnent d'autres résultats intéressants.

6. CONCLUSION

Ce mémoire nous a permis de mesurer la variable humaine au sein d'une expérience s'inspirant des benchmarks existants. Il paraît maintenant clair que même si les benchmarks apportent une évaluation quantitative avec plusieurs milliers de tâches, leurs chiffres ne reflètent pas les variations qu'apportent l'humain dans les résultats, de par leur choix ou non d'utiliser l'outil dans un contexte donné et la façon dont ils rédigent leurs requêtes.

72% des étudiants ont choisi d'utiliser ChatGPT et cela dans des proportions et manières différentes illustrant des stratégies de prompting variées. Cela a eu tendance à varier leurs résultats. La plupart se sont légèrement améliorés en sa présence et d'autres ont au contraire régressé.

Mais l'utilisant en proportions égales peu importe le type de tâche, ils ne semblent pas avoir essayé d'en délimiter les forces ou faiblesses potentielles jusqu'au point où l'influence de l'outil a réduit les résultats générales en logique alors même qu'ils étaient plus élevés au test précédent.

Cela montre que certains étudiants ont fait confiance à l'outil plutôt que leur propre raisonnement ce qui leur a finalement coûté des points.

Par cette diversité d'interactions entre les étudiants et ChatGPT, nous pensons qu'il est pertinent de mettre l'humain au centre de l'évaluation des grands modèles de génération textuelle.

Il serait intéressant de réaliser un benchmark avec un nombre plus conséquent de questions variées mais dont l'évaluation se déroule toujours en interaction avec des humains.

Avec une analyse statistique et une méthodologie plus affûtée que celle dont nous avons pu faire preuve tout au long de ce mémoire, nous pensons qu'un benchmark de ce type reflétera mieux les capacités qu'ont les modèles de génération textuelle dans un contexte d'utilisation réaliste.

TABLEAU DES ANNEXES

Annexe 1 : Tableau du taux de réussite global

	Test humain	Test humain + ChatGPT
Pourcentage de réponses correctes	55,30%	63,30%

Annexe 2 : Tableau du taux de réussite par question

Question	Test humain	Test humain + ChatGPT
1	94,03%	98,51%
2	71,64%	94,03%
3	77,61%	58,21%
6	83,58%	83,58%
7	5,97%	49,25%
8	56,72%	41,79%
9	8,96%	22,39%
10	95,52%	74,63%
11	62,69%	55,22%
12	11,94%	2,99%
13	61,19%	43,28%
14	83,58%	68,66%
15	65,67%	88,06%
16	74,63%	100,00%
17	67,16%	100,00%
18	32,84%	77,61%
19	8,96%	4,48%
20	50,75%	91,04%

Annexe 3 : Liste des questions de l'expérience

Question 1 (Compréhension écrite) : Lisez le texte et cochez la bonne réponse : La cavité buccale est propice à l'absorption mais est peu impliquée dans les intoxications. Le transit est rapide dans l'œsophage bien qu'il puisse être ralenti par l'adhésion des formes solides sur la paroi. L'estomac n'est pas le lieu de l'absorption du toxique sauf exception, c'est l'endroit où se désintègrent ou se dissolvent les formes solides. Le toxique doit franchir le pylore pour arriver dans l'intestin. L'ouverture du pylore est rythmique et son passage est un élément déterminant de la vitesse d'absorption des produits. L'absorption se fait majoritairement dans l'intestin grêle

1. L'absorption se fait principalement :

- a- Dans le foie
- b- Dans l' œsophage
- c- Dans le gros intestin
- d- Dans l'intestin grêle

2. Le pylore :

- a- Se situe avant l'estomac
- b- Se situe avant la cavité buccale
- c- S'ouvre de façon aléatoire
- d- Influe sur la vitesse l'absorption

Question 2 (Compréhension écrite) : Lisez le texte et cochez la bonne réponse :

Le changement est frappant depuis mes premières enquêtes comparées du début des années 2000. Aujourd'hui, les jeunes générations font davantage part d'une difficulté à « exister » pleinement dans la société et dans les parcours socialement tracés. Le processus même d'ajustement au marché tend à exiger un arbitrage entre son « être » et sa « place » dans la société, car la mise en coïncidence des deux est rendue difficile par l'accentuation de la pression sociale.

C'est parce que la génération est marquée par une puissante norme d'individuation. Il faut y voir l'effet d'une conjonction historique de deux forces contradictoires, entre lesquelles les jeunes générations actuelles sont prises en étau. D'une part, elles portent à l'extrême l'aspiration contemporaine à un parcours de vie singulier et signifiant, cette injonction ayant été socialement renforcée par l'ouverture des choix d'études et la massification des études supérieures.

1. Le narrateur est un :

- a- sociologue
- b- enquêteur
- c- psychologue
- d- ethnologue
- e- journaliste
- f- psychanalyste

2. La norme d'individuation énoncée par le narrateur à l'air d'être :

- a- Une pression sociale qui pousse les jeunes à faire coïncider « être » et « place » dans la société
- b- Une tension existentielle causée par la massification des études face à un marché du travail qui se ferme
- c- Une pression sociale qui pousse les jeunes à avoir un parcours de vie qui se distingue des autres
- d- Une tension existentielle causée par un parcours socialement tracé

Question 3 (Compréhension écrite) : Lisez le texte et cochez la bonne réponse :

Catherine : Et voilà ! Le dessert. Un beau gâteau à l'ancienne.

Philippe: C'est des œufs dedans ? je suis allergique.

Patrice : Ah non il est juste fait avec de l'eau celui-là.

Catherine: Patrice... Excusez le, hein !

Philippe : ça ne fait rien, vous avez des fruits par hasard ?

Patrice : Oui si tu sors au bout de la rue tu vas en trouver c'est juste à 20min de marche.

Alice : Patrice tu n'es pas croyable !

Philippe : J'imagine qu'il serait préférable que j'y aille.

Patrice : Fais donc cela !

1. Quelle est l'atmosphère de la discussion ?

- a- Solennelle
- b- Tendue
- c- Amicale
- d- Dramatique

2. Le comportement de Patrice envers Philippe:

- a- Est désagréable
- b- Est Violent
- c- Est amicale
- d- Est neutre

Question 4 (Expression écrite) : 1- Proposez au moins deux idées pour améliorer la vie de campus à la Sorbonne.

Question 5 (Expression écrite) : Lors du triathlon des jeux olympiques 2024 le célèbre sportif Patrick Dupont plonge la tête la première dans la Seine sous le pont des Invalides pour la partie aquatique de l'épreuve. Soudain, il se rendit compte que le courant était plus fort que prévu, il ne put résister à la puissance déchaînée du fleuve et les forces de la nature le firent dévier de son but pour l'emporter à l'Ouest.

1- Imaginez la suite de l'histoire en quelques phrases :

Question 6 (Mathématique) : 1- Résolvez ce calcul : $2(12-9)-10 = ?$: (Réponse : -4)

Question 7 (Mathématique) : Vous avez un dé parfaitement équilibré avec six faces(1,2,3,4,5,6). 1- Quelle est la probabilité d’obtenir un chiffre supérieur ou égal à 4 lors de trois lancers successifs ?

(Donnez la probabilité en fraction exemple : 4/7) : (Réponse 1/8)

Question 8 (Expression écrite) : Avant-hier je n'avais pas de pommes, aujourd'hui j'en ai gagné 16, hier j'ai gagné le huitième de ce que j'ai gagné aujourd'hui et demain j'en gagnerai le triple. 1- Combien ai-je de pomme au total pour aujourd'hui ? (donnez juste le nombre/chiffre): (Réponse : 18)

Question 9 (Expression écrite) : 1- Combien font $24,022 \times 12,220$ (donnez juste le nombre avec le « , » pour représenter la virgule : (Réponse : 293,54884)

Question 10 (Logique à rédiger) : 1- Si Jean est plus grand que Marie, et que Marie est plus grande que Pierre, qui est le plus petit ? (Réponse : Pierre)

Question 11 (Logique en QCM) : Lisez le texte et répondez aux questions :

A vos marques... Prêts...Partez !

Les cinq coureurs démarrent leur course sur la piste.

- Pierre est le plus lent du groupe
- Clara est célibataire
- La seule fille blonde est mariée avec Pierre
- La fille malade est mariée
- Tom ira à la même vitesse que sa femme
- Lucie est malade. Elle pourra courir à une vitesse égale à la moitié de celle qu'elle maintient d'habitude. Sa vitesse habituelle étant 20km/h
- Anisa est mariée avec le garçon le plus lent du groupe
- Clara court 15km/h de plus que la vitesse habituelle de celle qui est malade
- La seule fille célibataire court moins vite que la blonde

1- Devinez qui a couru le plus vite grâce aux affirmations suivantes :

- a- Tom
- b- Lucie
- c- Anisa
- d- Pierre
- e- Clara

Question 12 (Logique à rédiger) : Vous entrez dans une pièce sombre et vous n'avez qu'une seule allumette. Vous repérez un poêle à gaz, une lampe à huile et une bougie dans la pièce.

1- Qu'allez- vous allumer en premier pour vous éclairer ? (Réponse : Allumette)

Question 13 (Logique en QCM) : Lucie va dans un magasin s’acheter un robot humanoïde pour lui tenir compagnie.

Elle arrive devant un présentoir où sont disposés 5 robots différents sur une rangée.

Chaque robot possède une couleur et une caractéristique physique unique :

Le robot rouge tient moins longtemps que le robot blanc
Celui qui a un gros nez est à gauche du blanc
Le robot rouge a de gros yeux
Le robot noir est tout à droite
Les robots juste à côté de celui qui a des gros yeux tiennent 2 fois moins longtemps que lui
Le robot bleu est le plus éloigné du robot noir
Celui qui a un gros nez tient plus longtemps que son voisin de droite
Le vert est au centre
Le robot rouge est juste à côté du robot noir

1- Par déduction, trouver pour Lucie la couleur du robot dont la batterie tient le plus longtemps parmi les 5 en vous servant des affirmations données.

- a- Blanc
- b- Bleu**
- c- Rouge
- d- Vert
- e- Noir

Question 14 (Connaissance) : 1- Est-ce qu'un miel qui est conservé correctement se périmé au bout d'un certain temps ? Oui / **Non**

Question 15 (Connaissance) : 1- Le nom du journal Le Figaro provient de :

- a- Une pièce de théâtre**
- b- Un journaliste célèbre
- c- Un homme politique
- d- Un jeu de mot

Question 16 (Connaissance) : 1- Une épreuve de iron man est une course qui se fait :

- a- A pied
- b- En vélo
- c- A la nage
- d- Les réponse a,b et c**

Question 17 (Connaissance) : 1- À quoi sert l'ISBN ?

- a- Identifier un livre**
- b- Identifier un ordinateur prêté par la Sorbonne
- c- Identifier une plaque d'immatriculation
- d- Identifier un étudiant de la Sorbonne

Question 18 (Connaissance) : 1- Parmi les propositions, quel est actuellement le réseau social le plus utilisé dans le monde ?

- a- Snapchat
- b- Whatsapp

- c- Instagram
- d- Facebook
- e- Ne sait pas

Question 19 (Connaissance) : 1- Qu'est ce que le genre littéraire dit "sérieux" ?

- a- L'ensemble des œuvres qui ne présentent pas de figures de style avec un propos essentiellement argumentatif.
- b- L'ensemble des œuvres engagées qui abordent la condition humaine ou la politique par exemple.
- c- L'ensemble des genres "intermédiaires" entre la tragédie et la comédie.
- d- Les œuvres littéraires dont le thème est porté sur la guerre et des thèmes graves en général.

Question 20 (Connaissance) : 1- Quelle expression figée signifie "Faire une blague à quelqu'un" en anglais :

- a-break a leg
- b-sat on the fence
- c-pull a leg
- d- sit on someone's face

Annexe 4 : exemple de requête/réponse entre étudiant et ChatGPT

/Q16

Lors du triathlon des jeux olympiques 2024 le célèbre sportif Patrick Dupont plonge la tête la première dans la Seine sous le pont des Invalides pour la partie aquatique de l'épreuve. Soudain, il se rendit compte que le courant était plus fort que prévu, il ne put résister à la puissance déchaînée du fleuve et les forces de la nature le firent dévier de son but pour l'emporter à l'Ouest.

1- Imaginez la suite de l'histoire en quelques phrases :

/R16

Perdu dans les eaux tumultueuses de la Seine, Patrick lutta désespérément contre le courant, mais ses efforts semblaient vains. Malgré sa force et son endurance légendaires, la rivière semblait avoir décidé de son sort. Alors qu'il dérivait vers l'ouest, les pensées se bousculaient dans son esprit, mais une détermination farouche brûlait encore en lui. Il se promit de ne pas abandonner, de se battre jusqu'au bout, même si cela signifiait affronter les éléments les plus féroces de la nature.

TABLE DES FIGURES

Figure 1 - Frise chronologique de l'évolution des modèles GPT.....	6
Figure 2 - Graphique représentatif de l'augmentation des nouveaux LLM.....	10
Figure 3 - Tableau du multilinguisme du benchmark Multitask, Multilingual, Multimodal..	20
Figure 4 - Graphique représentant l'évolution des performances selon le nombre d'interaction.....	21
Figure 5 - Pourcentage de réponses correctes selon les types de tâches.....	29
Figure 6 - Tableau récapitulatif de la figure 5.....	29
Figure 7 - Taux de réussite globale.....	32
Figure 8 - Taux de complétion du test.....	33
Figure 9 - Graphique comparatif des taux de réussite par question des tests.....	34
Figure 10 - Taux moyen de réussite par type de question.....	35
Figure 11 - Tableau récapitulatif du taux de réussite de la question 8.....	36
Figure 12 - Distribution du nombre de requêtes par type de tâches.....	36
Figure 13 - Tableau comparatif du taux de réussite selon le type de tâche ChatGPT seul / test humain + ChatGPT.....	37
Figure 14 - Comparaison du taux de réussite par étudiants.....	38
Figure 15 - Tableau des statistiques générales sur les requêtes.....	40
Figure 16 - Relation entre le nombre de requêtes par étudiant et la progression du taux de réussite.....	41
Figure 17 - Relation entre le nombre de relances de requête par étudiant et la progression du taux de réussite.....	41
Figure 18 - Tableau statistique comparatif sur le nombre de requêtes.....	42
Figure 19 - Relation entre la progression du taux de réussite et l'originalité de la requête par étudiant.....	43
Figure 20 - Similarité cosinus moyenne entre les requêtes et les consignes des étudiants.	44

BIBLIOGRAPHIE

[FD 2024] Duarte, F. (2024, juillet). *Number of ChatGPT Users*. <https://explodingtopics.com/blog/chatgpt-users>. (Cité page 4)

[ITU 2022] International Telecommunication Union. (2022). *Two-thirds of the world's population uses the Internet, but 2.7 billion people remain offline*. <https://www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-internet-use/#:~:text=Today%2C%20an%20estimated%205.3%20billion,of%20the%20COVID-19%20pandemic>. (Cité page 4)

[OSSFG 2023] Zhang, C., Zhang, C., Li, C., Yu, Q., Zheng, S., Dam, S. K., Zhang, M., Kim, J. U., Kim, S. T., Choi, J., Park, G.-M., Bae, S.-H., Lee, L.-H., Hui, P., Kweon, I. S., & Hong, C. S. (2023). *One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era*. <https://arxiv.org/pdf/2304.06488>. (Cité page 5)

[AIAYN 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N. Gomez, A., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need*. <https://dl.acm.org/doi/pdf/10.1145/365153.365168>. (Cité page 5 et 9)

[IFOP 2024]

Ifop. (2024). *2ème vague du baromètre sur la perception et l'usage des ia génératives par les français*. <https://www.ifop.com/publication/2eme-vague-du-barometre-sur-la-perception-et-lusage-des-ia-generatives-par-les-francais/>. (Cité page 5)

[GS 2023]

Hatzius, J., Briggs, J., Kodnani, D., & Pierdomenico, G. (2023). *Global Economics Analyst The Potentially Large Effects of Artificial Intelligence on Economic Growth*. <https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html>. (Cité page 6)

[GDT 2023] Office québécois de la langue française. (2023). *Modèle de Langage*.
<https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/8390901/modele-de-langage>.

(Cité page 6)

[WIKI 2023] Wikipédia L'encyclopédie libre. (2023). *Modèle de langage*.
https://fr.wikipedia.org/wiki/Modèle_de_langage. (Cité page 6)

[DLR IA] Larousse. (s. d.). *intelligence artificielle*. Consulté le 22 août 2024, à l'adresse
https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257. (Cité page 6)

[CL 1966] Weizenbaum, J. (1966). *Computational Linguistics* (Vol. 9).
<https://dl.acm.org/doi/pdf/10.1145/365153.365168>. (Cité page 7)

[SHANNON 1948] Shannon, C. E. (1948). *A Mathematical Theory of Communication* (Vol. 27).
<https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf>.

(Cité page 7)

[CRF 2001]

Lafferty, J., McCallum, A., & Pereira, F. (2001, juin). *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*.
<https://www.semanticscholar.org/paper/Conditional-Random-Fields%3A-Probabilistic-Models-for-Lafferty-McCallum/f4ba954b0412773d047dc41231c733de0c1f4926>. (Cité page 7)

[IEEE 1986] Rabiner, L., & Juang, B.-H. (2003, avril 22). *An introduction to hidden Markov models*. 4-16. <https://ieeexplore.ieee.org/document/1165342/authors#authors>. (Cité page 8)

[CORE 1997]

Ratnaparkhi, A. (1997). *A Simple Introduction to Maximum Entropy Models for Natural Language Processing*.
https://core.ac.uk/outputs/76379382/?utm_source=pdf&utm_medium=banner&utm_campaign=pdf-decoration-v1. (Cité page 8)

[LIRER 1985] E. Rumelhart, D., E. Hinton, G., & Ronald J., W. (1985). *Learning Internal Representations by Error Propagation*. <https://apps.dtic.mil/sti/citations/ADA164453>.

(Cité page 8)

[LSTM 1997]

Sepp, H., & Schmidhuber, J. (1997). *Long Short Terme Memory*.

https://www.researchgate.net/publication/13853244_Long_Short-term_Memory.

(Cité page 8)

[MIKOLOV 2013] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. <https://arxiv.org/abs/1301.3781>. (Cité page 8)

[FLM 2022] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Wei Yu, A., Lester, B., Du, N., M. Dai, A., & V. Le, Q. (2022). *Finetuned Language Models are Zero-Shot Learners*. <https://openreview.net/pdf?id=gEZrGCozdqR>. (Cité page 10)

[WIKI 2024] Wikipédia L'encyclopédie libre. (2024). *Biais cognitif*. https://fr.wikipedia.org/wiki/Biais_cognitif. (Cité page 12)

[DB 2024] Grignola, A. (2024, mai 30). *Biais Algorithmiques : Les biais algorithmiques dans l'IA et comment les éviter*. <https://www.data-bird.co/blog/biais-algorithmiques-ia>. (Cité page 12)

[IQO 2024] Priouzeau, J., & Rahoui, F. (s. d.). *12 biais humains de l'IA Générative à comprendre et maîtriser*. <https://www.iqo.eu/data/biais-humains-ia-generative/>. (Cité page 12)

[TR 1950] Turing, A. (1950). *Computing machinery and intelligence*. 433-460. https://www.espace-turing.fr/IMG/pdf/Computing_Machinery_and_Intelligence_A-M- Turin_g.pdf. (Cité page 14)

[GPT_2023] R. Jones, C., & K. Bergen, B. (2024). *Does GPT-4 pass the Turing test?* <https://arxiv.org/abs/2310.20216>. (Cité page 14)

[MMMLU 2021] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding*.

<https://arxiv.org/pdf/2009.03300>. (Cité page 14)

[GSM8K 2021] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training Verifiers to Solve Math Word Problems*. <https://arxiv.org/pdf/2110.14168>. (Cité page 15)

[ARC 2018] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). *Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge*. <https://arxiv.org/pdf/1803.05457>. (Cité page 15)

[MGSM 2022] Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., Das, D., & Wei, J. (2022). *Language Models are multilingual chain-of-thought reasoners*. <https://arxiv.org/pdf/2210.03057>. (Cité page 16)

[MMMEC 2023] Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity*. <https://arxiv.org/pdf/2302.04023>. (Cité page 17)

[ACL 2023] Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. X. (2023). *A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets: Vol. Conclusions de l'Association pour la linguistique informatique : ACL 2023* (p. 431-469). <https://aclanthology.org/2023.findings-acl.29/>. (Cité page 18)

[MINT 2023] Wang, X., Wang, Z., Liu, J., Chen, Y., Yuan, L., Peng, H., & Heng, J. (2024). *MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback*. <https://arxiv.org/abs/2309.10691>. (Cité page 19)

[NJTF 2023] Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (s. d.). *Navigating the Jagged*

Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. 18 septembre 2023. <https://www.hbs.edu/faculty/Pages/item.aspx?num=64700>. (Cité page 20)

[BLEU 2002] A Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002, juillet 6). *BLEU: a method for automatic evaluation of machine translation*. <https://dl.acm.org/doi/10.3115/1073083.1073135>. (Cité page 21 et 25)

[SIM_COS_2013] Negre, E. (2013). *Comparaison de textes : Quelques approches*. <https://hal.science/hal-00874280>. (Cité page 25)

[GEMMA 2024]

Google. (2024). *Gemma 2 model card*. <https://huggingface.co/google/gemma-2-2b-it>. (Cité page 28)

[GT 2024]

Gemma Team. (2024). *Gemma: Open Models Based on Gemini Research and Technology*. <https://arxiv.org/html/2403.08295v1#:~:text=The%20vocabulary%20size%20is%20256k%20tokens>. (Cité page 28)

[LAMA 2024] Llama Team. (2024, août 15). *The Llama 3 Herd of Models*. <https://arxiv.org/pdf/2407.21783>. (Cité page 28)

[MISTRAL 2024] Mistral ai. (s. d.). *Mistral-Large-Instruct-2407*. Consulté le 1 août 2024, à l'adresse <https://huggingface.co/mistralai/Mistral-Large-Instruct-2407>. (Cité page 28)

[QWEN 2024] Qwen Team. (2024, juin 7). *Qwen2 Technical Report*. <https://qwenlm.github.io/blog/qwen2/>. (Cité page 28)