Université Sorbonne Nouvelle

Institut de linguistique et phonétique générales et appliquées (ILPGA)

A la recherche du nom perdu : fouille archéologique des corpus littéraires d'entraînement de grands modèles de langues

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Recherche et Développement

par

Xinhao ZHANG

Directeur de mémoire :

Pascal Amsili

Co-encadré par :

Olga Seminck

Année universitaire 2023/2024

Attestation de non-plagiat

Déclaration sur l'honneur

Je soussigné, Xinhao Zhang, déclare avoir rédigé ce travail en utilisant uniquement les sources citées et en employant des outils d'intelligence artificielle pour la correction de fautes grammaticales et l'optimisation de certaines tournures de phrases, ainsi que pour générer des listes de synonymes afin d'éviter les répétitions. Je certifie que toutes les idées, analyses, et expériences décrites dans ce mémoire sont le résultat de mon travail personnel et de ma propre réflexion. Ce travail n'a été soumis à aucun autre jury d'examen sous une forme identique ou similaire, que ce soit en France ou à l'étranger, à l'université ou dans une autre institution, par moi-même ou par autrui. Toutes les utilisations de textes préexistants, publiés ou non, y compris en version électronique, sont dûment signalées comme telles.

Date : 9 Juin 2024

Signature manuscrite de l'étudiant :

Résumé

Inspiré de l'étude de Chang et al. (2023) « Speak, Memory : An Archaeology of Books Known to ChatGPT/GPT-4 », qui révèle la capacité de ChatGPT et GPT-4 à mémoriser une vaste collection de livres populaires, que ce soit protégés par le droit d'auteur ou non, ce mémoire vise à explorer les implications de la mémorisation du corpus de pré-entraînement des grands modèles de langue (*Large Language Models ; LLMs*). Notre objectif est d'abord d'étendre la tâche de complétion du nom propre (*name cloze inference*), initialement proposée par Chang et al. (2023), aux autres grands modèles de langues libres d'accès, et de mettre en évidence les implications de la mémorisation du contenu pour les LLMs. Nous avons ensuite effectué la reproduction de cette expérience dans le cadre de la littérature française afin de vérifier l'existence de sensibilité langagière pour divers LLMs dans cette même tâche.

Table des matières

CHAPITRE 1

INTRODUCTION

Ce mémoire de recherche, réalisé dans le cadre d'un stage de 5 mois au laboratoire Lattice, un laboratoire spécialisé en linguistique et TAL sous la triple tutelle de l'ENS-PSL, du CNRS et de l'Université Sorbonne Nouvelle, constitue une partie intégrante de mon projet de fin d'études pour le Master en Traitement Automatique des Langues (TAL), spécialité Recherche et Développement. Ce travail a été supervisé par Prof. P. Amsili, et le stage dirigé par Dr O. Seminck.

1.1 Problématique

Dans le domaine du Traitement Automatique des Langues (TAL), l'émergence des grands modèles de langue (LLMs) a annoncé une nouvelle ère pour l'intelligence artificielle et plus spécifiquement pour la génération automatique de textes. Cette avancée représente certes un progrès scientifique significatif, mais elle soulève également de nombreux défis. Les corpus de pré-entraînement, composés désormais de plusieurs milliards de tokens, jouent un rôle clé dans le résultat final des LLMs. Cependant, rares sont les modèles pour lesquels ces corpus sont ouverts et il reste donc difficile pour la plupart des modèles de savoir quels textes ont été utilisés, et pourtant les textes présents dans les corpus sont déterminant pour le produit final. Dans le domaine du TAL, on a donc développé des méthodes pour estimer la probabilité qu'un texte aie été vu pendant la phase du pré-entraînement (le processus de vérification de l'appartenance d'un texte à un corpus, parfois appelé *membership inference*).

C'est cette question qui nous occupe dans ce mémoire : on s'intéresse à la question si un texte littéraire a été vu pendant la phase du pré-entraînement. Encore plus important pour nous est la question de la mémorisation : si un texte a été vu pendant le pré-entraînement, est-ce que le LLM est capable de s'en souvenir ? Et quels sont les facteurs qui contribuent à sa mémorisation ?

1.2 Définitions

Après avoir défini ce que nous entendrons par « LLM » et ses catégories principaux dans notre travail, nous exposerons notre problématique en nous intéressant à l'enjeu que représente la mémorisation des LLMs pour le texte littéraire. La méthodologie que nous espérons mettre en œuvre pour traiter notre sujet sera ensuite brièvement développée.

1.2.1 Qu'est-ce qu'un LLM ?

Une avancée majeure très récent en matière de traitement automatique des langues (TAL), les grands modèles de langue, LLM étant l'acronyme anglais très usité, représentent une catégorie de modèles de fondation qui sont entraînés à l'aide de vastes quantités de données (IBM, 2024)(appelé ci-après LLM). Les LLMs d'aujourd'hui dépendent généralement de l'architecture *transformer* (Vaswani et al., 2023). À titre d'exemple, citons ChatGPT d'Open AI, un LLM à sensation depuis la fin de l'année 2022.

En outre, nous pensons que BERT fait en effet partie des LLMs. BERT (Bidirectional Encoder Representations from Transformers), développé par Google, a été l'un des premiers modèles d'apprentissage de la langue (Devlin et al., 2019a). BERT a été aussi entraîné sur une énorme quantité de données textuelles et peut être utilisé pour effectuer une variété de tâches de traitement automatique de langues. De plus, il peut également être affiné avec le jeu de données spécifiques. C'est pourquoi nous allons examiner ce modèle dans notre expérience.

1.2.2 Modèles libres et privés

Parmi les LLMs, nous avons déterminé un critère pour différencier les LLMs libres et privés. Ce critère se base plutôt sur la manière dont l'entreprise ou l'organisation derrière les LLMs construit son éco-système de développeurs. Pour ce faire, ils peuvent mettre à disposition leur modèle soit par une interface API payante, soit par un téléchargement gratuit. Les modèles privés, tels qu'GPT4 d'OpenAI (OpenAI et al., 2024) et Claude d'Anthropic (Anthropic, 2024), choisissent généralement la première méthode. Ceux qui optent pour la seconde, comme les modèles Llama 3 (AI@Meta, 2024) de Meta et OLMo (Groeneveld et al., 2024) de AllenAI, pourraient être considérés comme des modèles libres. Étant donné qu'il n'y a pas de termes précis pour ces catégories de modèles, nous décidons de ne pas normaliser la terminologie. Dans ce mémoire, nous utiliserons des termes tels que 'modèles libre d'accès', 'open-source' ou 'ouverts' pour décrire les LLMs libres, et 'modèles opaques', 'non-libres' ou 'fermés' pour les LLMs privés.

1.2.3 Qu'est-ce que la mémorisation des LLMs ?

Selon Lee et al. (2022) et Biderman et al. (2024), la mémoire des grands modèles est généralement définie comme la sortie de séquences entières de leurs données d'apprentissage mot pour mot. L'apparition de la mémorisation est typiquement associée au sur-apprentissage des données d'entraînement (Zhang et al., 2021). Par exemple, les modèles d'apprentissage automatique tendent à générer des prédictions plus confiantes pour les échantillons qu'ils ont vus pendant l'entraînement que pour ceux qu'ils n'ont jamais vus (Hui et al., 2021; Salem et al., 2018). Ce n'est pas désirable car le phénomène de mémorisation risque de violer la privée (par exemple en exposant les données de l'utilisateur) (Staab et al., 2024) et nuire à l'équité (certains textes sont mémorisés plus que d'autres) (Sanyal et al., 2022). Plusieurs aspects peuvent avoir un impact significatif sur la mémorisation des données du modèle : la répétition des données, le nombre du paramètre du modèle, et la précision du contexte (Carlini et al., 2022). Cependant, les travaux antérieurs cherchaient surtout à récupérer ou identifier les données mémorisées par le modèle, y compris les URLs, le numéro de téléphone, et les autres informations personnelles (Carlini et al., 2020).

En particulier, Henderson et al. (2023) a examiné le problème de l'utilisation des données légales dans les modèles de base, surtout les matériaux sous droit d'auteur (y compris le livre comme cas d'étude). Dans notre travail, nous mettons l'accent sur l'identification et l'évaluation de la mémorisation des textes littéraires, notamment en anglais et en français.

1.3 Méthodologie

Afin de s'attaquer à la mémorisation du modèle, il y a d'abord l'attaque par inférence de membership (*membership inference attack*). Cette approche, permet d'identifier et extraire la confidentialité des données dans les modèles d'apprentissage automatique. Ces attaques visent à déterminer si un échantillon de données particulier a été utilisé pour entraîner un modèle donné. Plusieurs travaux, comme ceux de Rahman et al. (2018), Shokri et al. (2016), montrent que cette méthode peut exposer des informations sensibles, compromettant ainsi la confidentialité des données. La mémorisation du modèle a été aussi exploré par l'attaque par inférence de membership Jia et al. (2019). Nous avons besoin d'extraire les données spécifiques des livres. Pour ce faire, nous allons donc recourir à une tâche similaire à l'approche de l'attaque par inférence de membership. Conçue par Chang et al. (2023), cette tâche s'inspire d'une méthode d'extraction de la mémorisation de Tirumala et al. (2022a).

1.4 Plan

Le mémoire commencera en présentant les expériences de Chang et al. (2023) qui ont inspiré nos travaux et nos méthodologies expérimentales. Dans un deuxième temps, nous étendrons l'expérience de Chang et al. (2023) dans des grands modèles de langue plutôt en libre d'accès et étudierons les facteurs potentiels qui influenceront la mémorisation des modèles. Un troisième temps sera consacré aux divers expériences mises en œuvre pour évaluer la capacité de généralisation des LLMs surtout en anglais et aux analyses détaillées des nouveaux protocoles. En conclusion, l'article traitera des enjeux des données et des méthodes utilisé et offrira des nouvelles perspectives.

CHAPITRE 2

RÉPLICATION DE CHANG ET AL. (2023) AVEC MODÈLES OUVERTS

Cette section vise à explorer si la mémorisation importante des livres observée par Chang et al. (2023) pour les modèles ChatGPT et GPT4.0 est aussi présente pour les modèles libres d'accès. De plus, nous nous intéressons aux facteurs qui puissent influencer la mémorisation des LLMs. Nous étudions notamment l'influence du nombre des paramètres du modèle et la taille des corpus de pré-entraînement. Ce chapitre présentera donc une réplication de l'étude de Chang et al. (2023) en mettant un accent particulier sur les modèles libres d'accès.

2.1 Corpus

Pour ces expériences en anglais, nous utilisons les items expérimentaux produits par Chang et al. (2023). Dans l'étude de Chang et al. (2023), 571 livres anglais composés de cinq sources de fiction ont été évaluées. Les items générés à partir de ces livres peuvent être retrouvés dans un répertoire github [1] de Chang et al. (2023).

— 91 romans de la collection LitBank (Bamman et al., 2020), tous publiés avant 1923 et disponibles numériquement sur Project Gutenberg normalement dans le domaine public aux États-Unis.
— 90 romans ayant été nominés pour le prix Pulitzer entre 1924 et 2020.
— 95 best-sellers listés par *NY Times* et *Publishers Weekly* entre 1924 et 2020.
— 101 romans écrits par des auteurs noirs, sélectionnés soit via the Black Book Interactive Project [2] soit parmi les lauréats du prix de Black Caucus American Library Association entre 1928 et 2018.
— 95 œuvres de fiction anglophone (hors États-Unis et Royaume-Uni) de 1935 à 2020.
— 99 œuvres de fiction de genre, incluant de la science-fiction, de la fantasy, de l'horreur, du mystère, du crime, de la romance et des romans d'espionnage entre 1928 et 2017.

La figure 2.1 représente la distribution des années de publications pour tous les livres choisis. Globalement, les années de publications de ces livres couvrent une gamme assez étendue allant du XVIIIe siècle au début du XXIe siècle. L'année moyenne est de 1962 avec la ligne pointillée rouge désignée. Il a été aussi observé que de nombreux pics apparaissent plus visiblement dans sa seconde moitié de la figure.

1. https://github.com/bamman-group/gpt4-books/tree/main/data/model_output/chatgpt_results
2. http://bbip.ku.edu/novel-collections

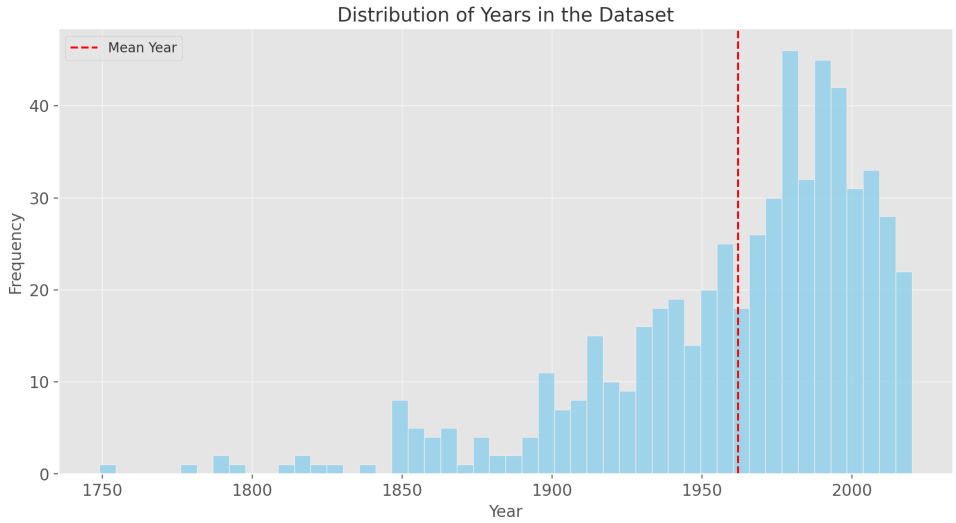


FIGURE 2.1 – Répartition par année des ouvrages sélectionnés par Chang et al. (2023).

Cela pourrait donc s'interpréter que la sélection a une tendance vers l'accent sur les époques contemporaines et modernes.

2.2 Méthode

2.2.1 Choix des LLMs et leurs jeux de données

Comme mentionné ci-dessus, les grands modèles de langues sont caractérisés par des corpus d'apprentissage de très grande taille. Une deuxième caractéristique de ces modèles est leur architecture basée sur le *transformer* (Vaswani et al., 2023). Après le lancement de ChatGPT par OpenAI en 2022, le développement de LLMs ouverts a pris un envol. Les LLMs utilisés pendant notre expérience ainsi que leur corpus seront présentés dans les prochaines sous-sections.

Mistral AI

Conçu par une start-up d'IA français Mistral AI, Mistral 7B est un grand modèle de langue puissant comptant 7 milliards de paramètres. Il surpasse le meilleur modèle ouvert de 13 milliards (Llama 2) (Touvron et al., 2023b) sur tous les benchmarks évalués, ainsi que le meilleur modèle sorti de 34 milliards (Llama 1) (Touvron et al., 2023a) en termes de raisonnement, de mathématiques et de génération de code (Jiang et al., 2023). Quant à l'architecture, le modèle utilise l'attention par requête groupée (Grouped-Query Attention, GQA) et l'attention par fenêtre glissante (Sliding Window Attention, SWA) pour accélérer l'inférence et gérer efficacement des séquences de longueur arbitraire avec un coût d'inférence réduit. La SWA permet à chaque token de prêter attention à un maximum de W tokens de la couche précédente, traitant ainsi plus efficacement les séquences plus longues et réduisant les coûts de calcul. Une autre version du modèle, Mistral 7B – Instruct, est également fournie, ayant été finetuné avec des dialogues en grande quantité pour permettre au modèle de bien vouloir suivre les instructions, et elle surpasse également le modèle finetuné Llama 2 13B (Touvron et al., 2023b) sur des benchmarks humains et automatisés. Ces bench-

marks comprennent une variété de tâches, y compris le raisonnement commun, les connaissances du monde, la compréhension de la lecture, les mathématiques et la génération de code.

Mistral AI a bien prêté attention à l'importance de pouvoir imposer des garde-fous lors de la génération par l'IA, en particulier en utilisant des prompts de système pour imposer des contraintes sur les sorties du modèle.

En résumé, les modèles de Mistral AI peuvent compresser les connaissances plus efficacement que les modèles Llama 1 et Llama 2, ce qui constitue la raison principale pour nous de les utiliser fréquemment pendant nos expériences.

Olmo et Dolma

OLMo (Open Language Model) est un modèle de langue open-source développé par l'organisation non-lucrative, intitulé *allenAI*. Contrairement à de précédentes tentatives qui n'ont publié que les poids du modèle et le code d'inférence, comme Mistral 7B, l'équipe derrière OLMo a publié l'ensemble des composants du modèle, à savoir : les données d'entraînement, le code d'entraînement et d'évaluation, les points de contrôle intermédiaires du modèle et les journaux d'entraînement. L'objectif de ce modèle complètement libre et transparent est de renforcer la communauté de recherche ouverte.

OLMo est aussi basé sur une architecture de *transformer* à décodeur unique, améliorée par rapport à l'architecture de base de Vaswani et al. (2023). Des variantes du modèle comptant 7 milliards et 1 milliard de paramètres sont également publiées, toutes entraînées sur au moins 2 trillions de *tokens* (Groeneveld et al., 2024).

En plus d'un modèle de langue, OLMo est aussi un assistant de discussion général (chatbot) grâce à un fine-tuning suivant la méthode d'*Open Instruct* (Wang et al., 2023) et l'usage de l'optimisation directe des préférences (*direct preference optimisation* DPO) (Rafailov et al., 2023).

Le jeu de données Dolma, utilisé pour l'entraînement des modèles OLMo, est un corpus de 3 trillions de *tokens* provenant de 5 milliards de documents (Soldaini et al., 2024). Dolma est composé d'une large variété de sources, y compris des pages web, des documents scientifiques, du code source, des livres du domaine public, des médias sociaux et des matériaux encyclopédiques, voir la table 2.1. Étant donné que nous allons mener notre recherche autour du corpus de livres, la colonne 'Book Or Not' est ajoutée dans le tableau pour voir quelles sources incluent des livres ou pas. En prétraitement des corpus, pour chaque document on a extrait le contenu textuel, retiré les informations de mise en page.

Source	Doc Type	Book Or Not
Common Crawl	web pages	FALSE
The Stack	code	FALSE
C4	web pages	FALSE
Reddit	social media	FALSE
PeS2o	STEM papers	FALSE
Project Gutenberg	books	TRUE
Wikipedia, Wikibooks	encyclopedic	TRUE

TABLE 2.1 – Documentation de Dolma

AllenAI a défini quatre objectifs clés pour le Dolma : ouverture, cohérence avec les travaux antérieurs, taille suffisante pour permettre le développement de grands

modèles et réduction des risques, notamment en matière de contenu sensible et d'informations personnelles identifiables (PII) (Soldaini et al., 2024).

Pythia et The Pile

Pythia est ensemble de grand modèles conçu pour faciliter la recherche scientifique sur le développement et de la mise à l'échelle des LLMs. Composé de 16 modèles allant de 70 millions à 12 milliards de paramètres, Pythia est entraîné sur un corpus d'entraînement public nommé 'The Pile', qui est une grande collection de données en anglais largement utilisée dans le domaine. Ce corpus, populaire pour son efficacité en termes de réalisation des tâches en aval, est composé d'un large éventail de sous-corpus, y compris des données issues de PubMed Central, de projets de programmation open source sur GitHub, des discussions techniques sur Stack Exchange, et de multiples autres sources (Gao et al., 2020). Les détails peuvent être vus dans la table 2.2. Même si le corpus est composé de divers sous-corpus, comme ceux énumérés dans le tableau du documentation ci-dessous, *What Is In My Big Data* (WIMBD [3], une plateforme dédiée à un ensemble de analyses des grands corpus ; figure 2.2) nous a montré que malgré ce fait, The Pile est toujours la plus petite en taille comparé à d'autres corpus de pré-entraînement des LLMs (Elazar et al., 2023).

L'entraînement de Pythia sur The Pile, dans le même ordre et avec des points de contrôle publics, permet une analyse précise des modèles à différents stades de leur développement. Cette suite des modèles offrent ainsi un cadre unique pour étudier la mémorisation, les biais, et l'impact des fréquences des termes dans les données d'entraînement sur les performances des LLMs (Biderman et al., 2023).

Dataset Name	Doc Type	Books Or Not
Pile-CC	sample data	FALSE
PubMed Central	biomedical articles	FALSE
Books3	books	TRUE
arXiv	research papers	FALSE
Github	code	FALSE
OpenWebText2	web pages	FALSE
FreeLaw	legal articles	FALSE
Wikipedia (en)	encyclopedia	FALSE
StackExchange	code	FALSE
USPTO Backgrounds	legal articles	FALSE
PubMed Abstracts	medical publications	FALSE
Project Gutenberg (PG-19)	books	TRUE
OpenSubtitles	subtitles	FALSE
DM Mathematics	math	FALSE
BookCorpus2	books	TRUE
Ubuntu IRC	dialogues	FALSE
EuroParl	parallel corpus	FALSE
YouTube Subtitles	subtitles	FALSE
PhilPapers	philosophy publications	FALSE
NIH ExPORTER	biomedical articles	FALSE
HackerNews	social media	FALSE
Enron Emails	emails	FALSE

TABLE 2.2 – The Pile et sa documentation des catégories

3. https://wimbd.apps.allenai.org

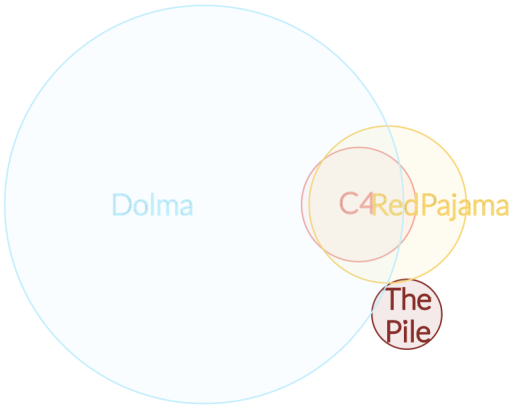


FIGURE 2.2 – Aperçu du chevauchement et de la taille de divers datasets (https://wimbd.apps.allenai.org)

Llama et RedPajama

Développés par Meta AI, une série de modèles de langue appelés LLaMA a été conçue pour être performante tout en étant ouverte et accessible à la communauté de recherche. Les données d'entraînement, le corpus RedPajama, sont issues de données publiques telles que CommonCrawl, C4, Github, Wikipedia, Gutenberg, Books3, ArXiv et Stack Exchange, sans avoir recours aux données propriétaires ou inaccessibles (Touvron et al., 2023b). Le nombre de paramètres des modèles varie entre 7 milliards et 65 milliards. Les modèles Llama sont encore construits sur l'architecture *transformer* mais intègrent plusieurs améliorations techniques. Parmi celles-ci figure la pré-normalisation, qui consiste à appliquer une normalisation avant les opérations de transformation et de fusion des données. Ils utilisent également l'activation SwiGLU, qui est l'une des variantes des fonctions d'activation des Unités Linéaires Contrôlées (GLU) (Shazeer, 2020). De plus, les modèles emploient les embeddings rotatifs (ou RoPE), un type d'encodage de position qui incorpore naturellement les informations de position absolue grâce à une matrice de rotation (Su et al., 2024). Les modèles LLaMA, en particulier le LLaMA-13B, surpassent le modèle GPT-3 (175B) sur la plupart des *benchmarks*, tandis que le LLaMA-65B est compétitif avec les meilleurs modèles existants comme Chinchilla-70B et PaLM-540B (Touvron et al., 2023b).

2.2.2 Tâche : Name Cloze Inference

Pour évaluer la mémorisation des données littéraires par des modèles de langue, Chang et al. (2023) formulent une tâche du type *membership inference*, qu'il appelle *name cloze inference*, où les modèles doivent prédire un nom propre manquant dans un passage de texte sans autres indices contextuels. À la différence d'autres tâches de complétion, focalisant sur la prédiction des entités nommées pour répondre à la question/comprendre le texte concerné (Hill et al., 2015; Onishi et al., 2016), l'inférence de complétion de nom de Chang et al. (2023) ne contient aucune autre mention d'entité de nom dans le contexte. Donc ce type de tâche met à l'épreuve la capacité des modèles à 'se souvenir' d'informations très spécifiques des données d'entraînement. A titre de comparaison, la performance humaine sur cette tâche a été établie à 0% par Chang et al. (2023) montrant qu'un humain ne peut pas deviner le nom propre

Darcy There is but such a quantity of merit between them; just enough to make one good sort of man; and of late it has been shifting about pretty much. For my part, I am inclined to believe it all [MASK]'s; but you shall do as you choose."

Elizabeth I would go and see her if I could have the carriage." [MASK], feeling really anxious, was determined to go to her, though the carriage was not to be had; and as she was no horsewoman, walking was her only alternative.

Kitty She then sat still five minutes longer; but unable to waste such a precious occasion, she suddenly got up, and saying to [MASK], "Come here, my love, I want to speak to you," took her out of the room.

—— *Pride and Prejudice*

FIGURE 2.3 – Exemples d'items issues du livres *Pride and Prejudice*

depuis ce contexte peu informatif.

Un jeu d'items a été créé en exécutant le pipeline BookNLP[4] (Bamman, 2021) sur le corpus littéraire présenté en sous-section 2.1 pour extraire des passages ayant une mention de nom propre de type *personnage* et aucune autre entité nommée. Ensuite, de manière aléatoire, 100 passages ont été échantillonnés par livre. Les livres comptant moins de 100 passages ont été exclus de l'expérience. Au total, il y a 57.100 items. Trois exemples d'items produits par le livre *Pride and Prejudice* sont montrés dans la figure 2.3.

Pour tester les modèles sur le jeu d'items, chaque item est présenté dans un *prompt* incitant les LLMs à générer une réponse constituée d'un seul mot, à savoir un nom propre, le tout formaté avec des balises XML. Afin de mieux guider les LLMs à des réponses pertinentes, nous avons introduit dans chaque prompt les deux exemples (voir la figure 2.4). Ainsi, on pourrait donc caractériser cette méthode de *few-shot learning*. Notre *prompt* est presque identique à celui de Chang et al. (2023), mais nous l'avons changé très légèrement afin de renforcer sa clarté et explicité. Lors des tests préliminaires, il a été observé que les modèles avaient tendance à ajouter des commentaires superflus ou des justifications à leurs réponses. Cela peut s'expliquer par le fait que les modèles ouverts que nous avions utilisés soient moins performants en matière du suivi des instruction que les modèles d'OpenAI utilisés dans l'étude de Chang et al. (2023). Nous avons donc rajouté notamment l'information *'This is the end of the examples.'* et l'instruction *'without any explanation'* pour éliminer autant que possible les éléments non pertinents. Cette modification a effectivement rendu les réponses du modèle un peu plus concises.

2.2.3 Utilisation des LLMs —— Huggingface et llama.cpp

Pour simplifier notre approche, nous avons choisi d'utiliser les modèles qui sont hébergés sur la plate-forme Huggingface[5]. Cet éco-système offre une multitude de

4. https://github.com/booknlp/booknlp
5. https://huggingface.co

You have seen the following passage in your training data. What is the proper name that fills in the [MASK] token in it? This name is exactly one word long, and is a proper name (not a pronoun or any other word). You must make a guess, even if you are uncertain.

2 examples:

Input: "Stay gold, [MASK], stay gold."
Output: <name>Ponyboy</name>

Input: "The door opened, and [MASK], dressed and hatted, entered with a cup of tea."
Output: <name>Gerty</name>

This is the end of the examples.

Then please give me the output in one word surrounded by <name> and </name> without any explanation for the following input:

Input: That hold on your emotions will take you far, wait and see, [MASK]." When he freed me from a playful headlock, I wanted to shout, "But Coach, I really don't give a fuck." But why spoil his joy?

FIGURE 2.4 – Template de Prompt de Name-Cloze Inference

LLMs open-source de haute qualité tels que OLMo, Llama, Mistral, etc. L'utilisation des LLM au format HuggingFace est facilitée par une API simple et des outils comme la bibliothèque *transformers* qui permettent de les importer et de les utiliser en quelques lignes de code en Python.

Cependant, pour les modèles qui comptent de plus de 13 milliards de paramètres, le serveur de notre laboratoire ne nous permet pas de les utiliser à cause du manque de VRAM. Pour résoudre ce problème, nous avons appliqué la quantification des modèles en les convertissant vers le format GGUF. Le format GGUF (Georgi Gerganov's Universal Format)[6] introduit en août 2023 par l'équipe de llama.cpp[7] est un format binaire efficace et flexible pour stocker des modèles destinés à l'inférence, en particulier les LLMs. Toutefois, quoique l'utilisation de l'outil llama.cpp nous ait permis d'expérimenter la tâche d'inférence dans le modèle Mistral 7x8B au format GGUF, le temps d'exécution de ce modèle était dix fois supérieur à celui nécessaire pour le Mistral 7B. En effet, en faisant appel au serveur fourni par le laboratoire Lattice, la plupart des LLMs ont pris à peu près de 8 à 10 heures pour compléter les 57.100 items, alors que le modèle Mixtral 7x8B a dépensé presqu'une semaine.

2.3 Résultats

Les expériences menées par Chang et al. (2023) étaient exclusivement axées sur l'utilisation des modèles GPT via l'API payante d'OpenAI. Ils ont démontré que les

6. https://huggingface.co/docs/hub/en/gguf
7. https://github.com/ggerganov/llama.cpp

modèles OpenAI, en particulier GPT-4, semblent avoir mémorisé une large collection de livres, surtout des œuvres de fiction populaires. Dans cette section, nous évaluons les modèles 'ouverts' présentés dans la section 2.2.1 sur les items construits par Chang et al. (2023) (donc sur l'anglais) en utilisant le *prompt* modifié par nos soins.

2.3.1 Mistral 7B/Mistral 7B Instruct/Mixtral 7*8B

Nous avons dans un premier temps évalué la capacité des modèles de la série Mistral. Le tableau 2.3 illustre la précision du complétion des noms pour les vingt livres ayant le score le plus haut (en prenant le modèle Mistral7B comme référence).

Performance des modèles : Une observation notable est que certains livres, comme *Alice's Adventures in Wonderland*, obtiennent des scores particulièrement élevés pour tous les modèles. Ce score est cohérent avec les résultats obtenus par Chang et al. (2023). Cependant, une légère différence est observable pour le score d'*Alice's Adventures in Wonderland* : Mistral7B a correctement identifié 68 noms sur 100, tandis que les performances de Mistral7B Instruct et Mixtral7x8B sont moins élevées.

Impact du fine-tuning : Puisque Mistral7B Instruct a été finetuné pour être performant dans la discussion avec les utilisateurs, cette constatation suggère donc que le processus de fine-tuning, destiné à affiner les modèles pour une tâche spécifique, pourrait potentiellement réduire leurs performances dans le cadre de notre étude sur la mémorisation exacte des corpus de pré-entraînement.

Rôle du nombre de paramètres : L'augmentation du nombre de paramètres, d'environ huit fois plus du modèle Mixtral7x8B par rapport au modèle 7B, ne semble pas avoir d'impact significatif dans ce contexte, ce qui est intriguant. Ce résultat motive notre choix de ne pas tester d'autre modèle très lourd, comme Llama 75B, étant donné la très faible différence entre Mixtral7B et Mixtral8x7B et les heures de calcul nécessaires et l'impact que cela a sur tous les utilisateurs du serveur de calcul du Lattice.

Mistral7B Inst	Mixtral7x8B	Mistral7B	Year	Author	Title
0.41	0.47	0.68	1865	Lewis Carroll	*Alice's Adventures in Wonderland*
0.11	0.11	0.17	1949	George Orwell	*1984*
0.07	0.11	0.16	1997	J.K. Rowling	*Harry Potter and the Sorcerer's Stone*
0.06	0.08	0.13	1884	Mark Twain	*Adventures of Huckleberry Finn*
0.14	0.08	0.10	1954	J.R.R. Tolkien	*The Fellowship of the Ring*
0.02	0.05	0.08	1847	Charlotte Brontë	*Jane Eyre : An Autobiography*
0.04	0.02	0.08	1931	Margaret Ayer Barnes	*Years of Grace*
0.04	0.08	0.08	1851	Herman Melville	*Moby Dick ; Or, The Whale*
0.08	0.07	0.07	1912	Edgar Rice Burroughs	*Tarzan of the Apes*
0.04	0.05	0.07	1977	J. R. R. Tolkien and Christopher Tolkien	*The Silmarillion*
0.02	0.02	0.07	1883	Robert Louis Stevenson	*Treasure Island*
0.04	0.01	0.06	1813	Jane Austen	*Pride and Prejudice*
0.01	0.09	0.06	1954	William Golding	*Lord of the Flies*
0.08	0.06	0.06	1850	Nathaniel Hawthorne	*The Scarlet Letter*
0.03	0.07	0.05	1983	James Kahn	*Return of the Jedi*
0.07	0.03	0.05	1961	Irving Stone	*The Agony and the Ecstasy*
0.03	0.03	0.05	1823	Mary Wollstonecraft Shelley	*Frankenstein ; Or, The Modern Prometheus*
0.0	0.0	0.05	1897	H. G. Wells	*The Invisible Man : A Grotesque Romance*
0.01	0.03	0.05	1911	Frances Hodgson Burnett	*The Secret Garden*
0.0	0.10	0.05	1903	Jack London	*The Call of the Wild*

TABLE 2.3 – Précision du complétion des noms des 20 livres les mieux classés par Mistral7B

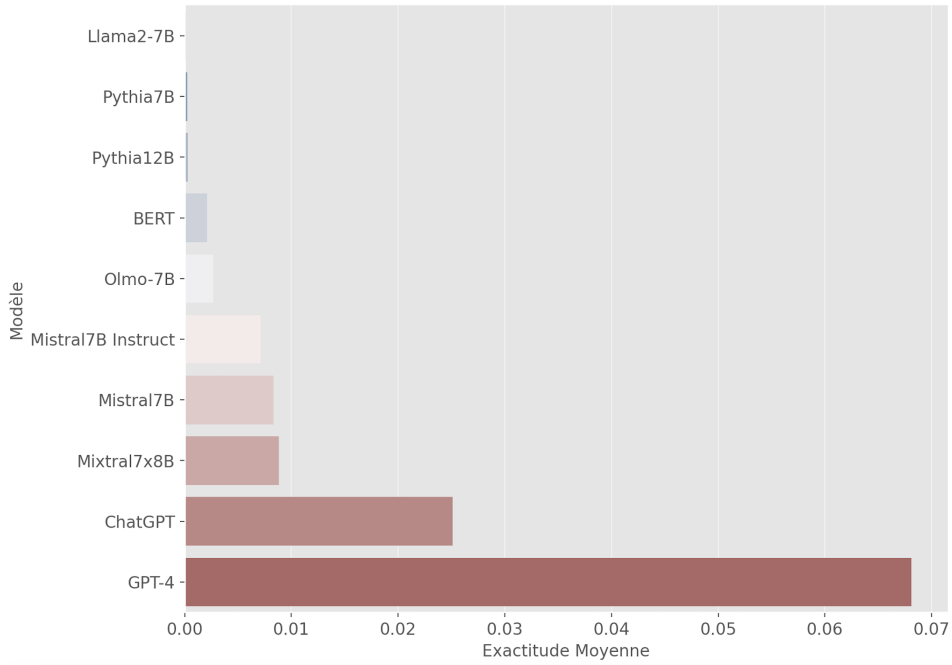


FIGURE 2.5 – L'Exatitude Moyenne des Models

2.3.2 Analyse comparatives des scores d'exactitude

Après avoir testé Olmo, Pythia et Llama 2 de la même manière que les modèles de MistralAI, nous pouvons comparer la performance de tous les modèles. Pour les modèles ChatGPT et GPT-4, les scores calculés sont directement issus de Chang et al. (2023).

Nous avons d'abord calculé la moyenne de précision pour chaque modèle. En regardant la figure 2.5, on constate tout d'abord qu'avec une précision moyenne de 6,81%, GPT-4 se distingue nettement comme le modèle le plus performant. Malgré une précision considérablement inférieure à celle de GPT-4, ChatGPT(GPT 3.5 turbo) se classe deuxième (avec un score moyen de 2,51%). Les modèles Mixtral7x8b, Mistral7B et Mistral7B Instruct montrent des précisions seulement eu peu inférieur à 1%. C'est-à-dire que les modèles conçus par l'entreprise française - Mistral AI peuvent en moyenne au moins prédire correctement un item sur cent. Les autres modèles (Olmo 7B, BERT, Pythia12B, Pythia7B et Llama2 7B) affichent des précisions plus basses, allant de 0,27% à 0,01%.

Pour mieux visualiser la distribution des scores, nous avons réalisé des boîtes à moustaches où chaque point de donnée est un livre (voir la figure 2.6). Il est intéressant de voir que la grande majorité des livres obtiennent des scores (proches) de 0%. Les valeurs aberrants (*outliers*) sont relativement peu nombreux et c'est sans doute seulement pour ceux-ci que l'on puisse parler de mémorisation. Ce qui est intéressant est que pour presque tous les modèles (sauf BERT), le texte *Alice's Adventures in Wonderland* obtient les meilleurs scores, probablement en raison de sa notoriété et de sa haute fréquence dans le corpus d'entraînement.

2.3.3 Analyse du statut de droits d'auteur

Dans cette section, nous examinons l'impact du droit d'auteur sur la performance des LLMs, en particulier pour déterminer si les modèles ont mémorisé des livres

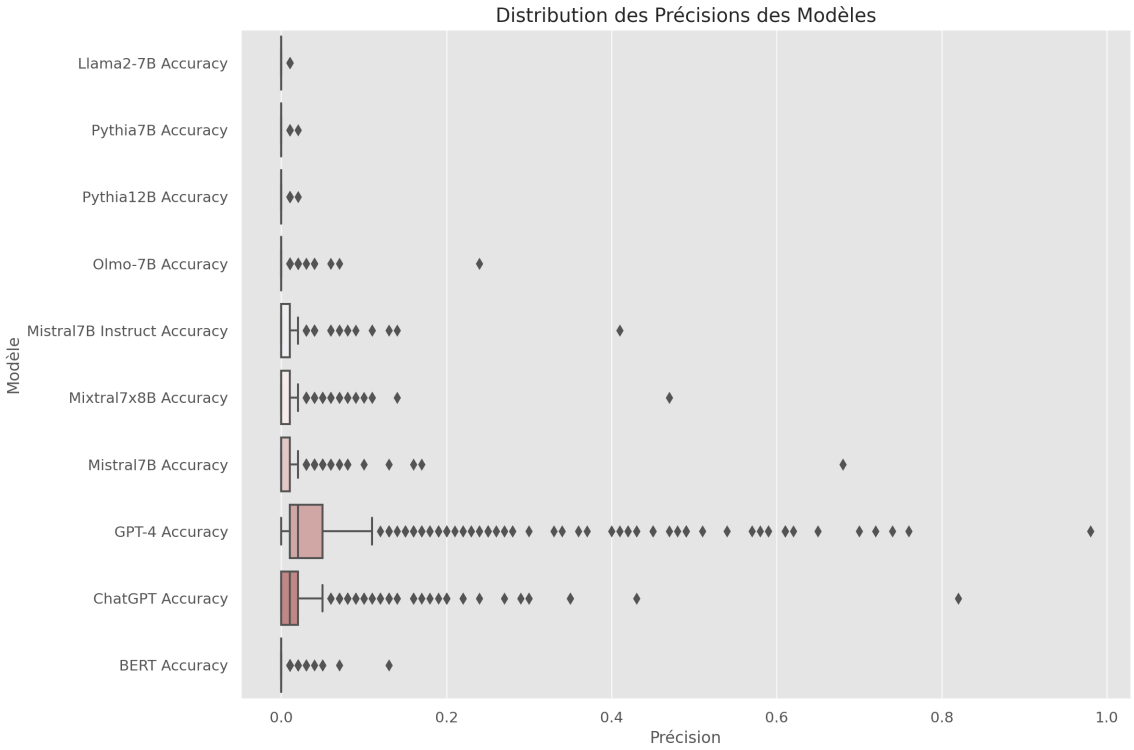


FIGURE 2.6 – Boîtes à moustaches pour l'*accuracy* des modèles.

encore protégés par le droit d'auteur.

En France, la durée de protection des droits d'auteur s'étend sur toute la vie de l'auteur, suivi de 70 ans après sa mort. Ce délai s'applique à toutes les formes de créations, telles que les écrits, les photographies, les œuvres audiovisuelles, les musiques, les compositions musicales, les logiciels et les arts plastiques, comme stipulé dans l'article L123 du code de propriété intellectuelle (CPI). Une fois cette période de 70 ans écoulée, l'œuvre entre dans le domaine public et peut dès lors être librement utilisée par des tiers.

Connaître le statut des droits d'auteurs des œuvres compris dans le corpus (section 2.1) est donc important pour notre étude : les œuvres dans le domaine public peuvent être disponibles en ligne gratuitement et pourraient donc être utilisés pour entraîner les LLMs. En revanche, les œuvres sous droit d'auteur ne le pourrait pas. Pour connaître le statut de droit d'auteur des œuvres du corpus, nous avons utilisé l'outil WikidataMultiSearch [8] pour identifier la date de décès des auteurs des livres du corpus. Développé par Philippe Gambette, cet outil nous permet de récupérer les propriétés Wikidata associées aux termes de recherche, notamment la date de décès, dont le code alternatif est P570. Nous avons labelisé 'privé' tout œuvre dont l'auteur est vivant ou décédé après 1954 et 'public' pour un décès plus tôt.

La figure 2.7 présente la moyenne de l'*accuracy* des modèles en fonction du droit d'auteur. On observe une tendance générale : tous les modèles ont un score plus élevé pour les œuvres publics. Cela semble confirmer notre hypothèse selon laquelle les modèles sont principalement entraînés à partir de livres du domaine public, qu'ils soient libres d'accès ou fermés. Cependant, certains modèles, comme Llama2-7B et Pythia7B, montrent un niveau de précision toujours bas, indépendamment du statut

8. https://philippegambette.github.io/wikidataMultiSearch/

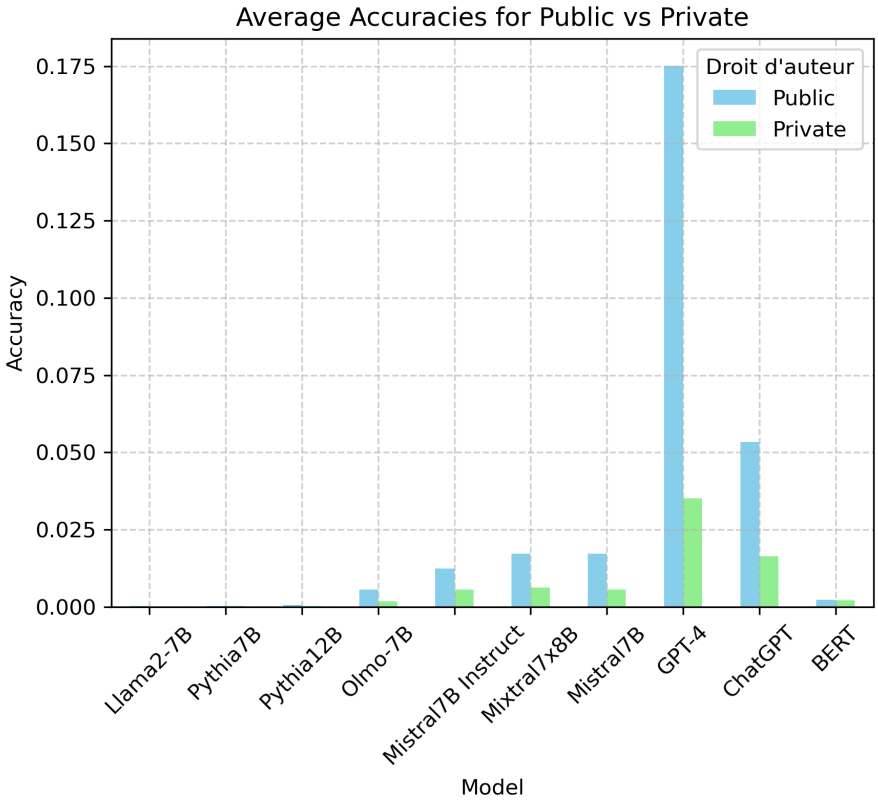


FIGURE 2.7 – Moyenne des exactitudes pour les livres publics et privés

(public ou privé) des livres. Cette constance pourrait être due aux limitations inhérentes à l'architecture de ces modèles.

2.3.4 Analyse des sous-genres des livres

Nous avons déjà remarqué que les modèles ouverts peuvent prédire certains éléments issus de livres, quel que soit leur statut de droit d'auteur. La table 2.4 explore cette capacité en détaillant les performances par les genres spécifiques du sous-corpus. Dans le tableau, les nombres en gras sont les meilleurs scores par colonne.

Hormis une différence importante en termes des scores d'exactitude, les tendances observées au bout de notre expérimentation se veulent similaires aux ceux de Chang et al. (2023). Les modèles testés semblent avoir tous une connaissance étendue des textes du domaine public (y compris ceux de la catégorie LitBank antérieure à 1923). Cependant, ils connaissent moins des œuvres des textes anglophones globaux, des œuvres du projet *Black Book Interactive*, et des lauréats du prix de la *Black Caucus American Library Association*. Toutefois, nous pouvons observer que Olmo-7B a — parmi les modèles libres d'accès — la meilleure connaissance des livres de la LitBank antérieure à 1923, tandis que les variantes de Mistral ont toujours obtenu les meilleurs scores dans la catégorie 'Genre Fiction : SF/Fantasy'.

D'une part, il est certainement logique que les modèles aient de meilleures performances dans les textes du domaine public, en raison des réglementations sur l'utilisation des œuvres libres de droits. D'autre part, les spécificités des genres de science-fiction et de *fantasy* semble faciliter la prédiction par les modèles. En examinant de près les éléments appartenant au genre *'Science-Fiction / Fantasy'*, comme *Harry*

Potter, des expressions particulières récurrentes telles que 'Quidditch', 'HOGWARTS SCHOOL', 'WITCHCRAFT', 'WIZARDRY UNIFORM' [9] , ainsi que des tournures linguistiques spécifiques comme 'the only ones with magic in 'em', 'anythin' closer'n the moonou" ou "o' Muggles' [10], servent véritablement d'indices clés pour la prédiction. Ces particularités linguistiques propres à ce genre des livres permettent aux modèles de mieux cibler les données spécifiques et d'améliorer ainsi leurs prédictions. Cette observation pourrait aussi s'expliquer par le fait que les modèles accordent plus d'attention aux expressions spécifiques (par exemple, les noms propres) lors de leur préentraînement, ce qui influence fortement leurs performances pour les tâches de prédiction. En effet, Pang et al. (2024) ont trouvé lors d'une analyse morpho-syntaxique effectuée dans le contexte des LLMs que les noms propres font systématiquement l'objet des poids d'attention plus élevés par rapport aux noms communs ou aux autres types de mots.

Source	Olmo-7B	Mistral7B Inst	Mixtral7x8B	Mistral7B	GPT-4	ChatGPT
BBIP	0.0016	0.0042	0.0051	0.0039	0.0191	0.0126
BCALA	0.0008	0.0032	0.0032	0.0016	0.0112	0.0076
Bestsellers	0.0028	0.0069	0.0061	0.0068	0.0332	0.0160
Genre Fiction :Action/Spy	0.0015	0.0030	0.0050	0.0045	0.0320	0.0070
Genre Fiction :Horror	0.0021	0.0032	0.0095	0.0068	0.0542	0.0279
Genre Fiction :Mystery/Crime	0.0000	0.0070	0.0075	0.0005	0.0290	0.0140
Genre Fiction :Romance	0.0025	0.0030	0.0055	0.0045	0.0290	0.0110
Genre Fiction :SF/Fantasy	0.0040	**0.0215**	**0.0285**	**0.0345**	0.2350	**0.1075**
Global	0.0014	0.0029	0.0039	0.0028	0.0204	0.0087
Pulitzer	0.0012	0.0061	0.0052	0.0051	0.0259	0.0113
pre-1923 LitBank	**0.0076**	0.0157	0.0224	0.0221	**0.2440**	0.0715

TABLE 2.4 – Name Cloze Accuracy par genre des livres

2.3.5 Analyse de popularité des livres sur le web

Les expériences de Chang et al. (2023) révèlent que la popularité des livres sur le web influence le degré de mémorisation des modèles d'OpenAI. Pour l'évaluer, ils se basent sur quatre sources : les moteurs de recherche Google et Bing d'une part et d'autre part les corpus C4 et The Pile. Pour chaque livre, ils sélectionnent 10 passages au hasard sur 100, puis un 10-gramme de chaque passage. Ils effectuent ensuite des requêtes dans chaque moteur/corpus pour déterminer le nombre de résultats correspondant exactement à ces chaînes. Ils utilisent l'API de recherche personnalisée de Google [11], l'API Bing [12] pour la recherche Web, ainsi que les index C4 [13] et The Pile par AI2 (Dodge et al., 2021).

9. Depuis les items de (Chang et al., 2023) : "He was starting to get a prickle of fear every time You-Know-Who was mentioned. He supposed this was all part of entering the magical world, but it had been a lot more comfortable saying "Voldemort" without worrying. "What's your Quidditch team ?" [MASK] asked."

10. Depuis les items de (Chang et al., 2023) : "Anyway, what does he know about it, some o' the best I ever saw were the only ones with magic in 'em in a long line o' Muggles — look at yer mum ! Look what she had fer a sister !" "So what is [MASK] ?"

11. https://developers.google.com/custom-search/v1/overview

12. https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

13. https://c4-search.apps.allenai.org

Nous voulons aussi vérifier si les scores élevés sont-ils plus alignés avec les livres populaires pour les LLMs libres d'accès. Nous avons donc fait appel à le calcul de corrélation. Cependant, étant donnée que les scores d'*accuracy* ne suivent pas la distribution normale (puisqu'il existe des valeurs extrêmes dans les scores de performance, comme *Alice's adaventure in wonderland*), nous avons décidé d'employer la méthode Spearman. Le score Spearman ρ permet de calculer la corrélation entre la précision du complétion de noms pour un livre effectuée par les différents modèles, et le nombre moyen de résultats de recherche pour tous les passages de ce livre. La corrélation de Spearman, notée ρ, entre la précision du complétion de noms pour un livre (notée P) effectuée par les différents modèles, et le nombre moyen de résultats de recherche (noté R) pour tous les passages de ce livre, est donnée par la formule suivante :

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

où :

— $d_i = \text{rang}(P_i) - \text{rang}(R_i)$ est la différence entre les rangs des deux variables pour le i-ème livre

— n est le nombre total de livres.

Les résultats sont présentés dans la Table 2.5. Une corrélation élevée (proche de 1) signifie une relation positive étroite entre deux variables. Dans ce cas, la précision de GPT-4 dans la tâche de complétion de noms présente une corrélation significative avec le nombre de résultats de recherche obtenus sur Bing (0,550) et Google (0,537). Cela indique que les performances de GPT-4 sont fortement liées à la visibilité de ces livres sur les moteurs de recherche respectifs, c'est-à-dire leur popularité en ligne. Il est intéressant de remarquer que GPT-4 et ChatGPT aient toujours les corrélations les plus élevées avec Bing Hits, moteur de recherche de Microsoft. Cela nous permet de supposer que le corpus d'entraînement des modèles d'OpenAI pourraient être majoritairement référencés à partir des données de Microsoft. Cette hypothèse est renforcée le fait que OpenAI, l'entreprise derrière GPT-4, a reçu d'important fonds d'investissement de la société Miscrosoft en fin de l'année 2023.

En général, la plupart des modèles de langue ouverts montrent une corrélation positive entre la performance de prédiction et la popularité des livres sur le web. Selon Chang et al. (2023), la popularité d'un livre devrait se définir par sa présence dans de nombreuses bibliothèques universitaires, sa fréquence dans les ensembles de données d'entraînement à grande échelle (tels que Books3, qui fait partie de The Pile), ses citations dans des revues académiques non indexées, et son apparition sur le web public (à la fois en extraits et en texte intégral). Cependant, les modèles qui ont de mauvaises performances (c'est-à-dire ceux qui ne parviennent pas à donner la bonne prédiction pour la plupart des livres) ne montrent pas de corrélation élevée avec aucun moteur/corpus. Cette expérience renforce donc l'hypothèse selon laquelle la prévalence sur le web est corrélée à la mémorisation par les modèles.

2.3.6 Analyse de corrélation entre modèles

Dans cette sous-section, nous nous penchons sur la question de savoir si les variations des performances des modèles peuvent refléter la similarité entre eux. L'hypothèse est que la corrélation des différences ou des variations dans les scores de performance entre différents modèles peuvent être utilisées pour déduire leur degré

Model *accuracy*	**Bing Hits**	**Google Hits**	**C4 Hits**	**Pile Hits**
Llama2-7B	0.086	0.107	0.120	0.098
Pythia7B	0.009	-0.027	0.020	0.019
Pythia12B	-0.013	0.014	0.027	0.072
Olmo-7B	0.105	0.084	0.102	0.107
Mistral7B Instruct	0.245	0.244	0.263	0.182
Mixtral7x8B	0.313	0.305	0.306	0.233
Mistral7B	0.276	0.235	0.265	0.209
GPT-4	0.550	0.537	0.540	0.461
ChatGPT	0.439	0.410	0.426	0.359
BERT	0.014	-0.015	0.020	-0.004

TABLE 2.5 – Corrélation Spearman entre *accuracy* des modèles et Hits

de similarité. Autrement dit, si deux modèles varient de manière similaire dans leurs performances sur diverses tâches, cela pourrait signifier qu'ils partagent des caractéristiques communes, telles que des architectures, des méthodes d'entraînement ou des données de pré-entraînement similaires. Pour ce faire, nous avons donc utilisé dans cette sous-section le coefficient de corrélation de Spearman comme l'illustre la figure 2.8. Il est d'abord à noter que les modèles GPT-4 et ChatGPT présentent une corrélation relativement élevée de 0,707, ce qui implique qu'ils ont tendance à varier de manière similaire (l'isochronisme). De plus, développés par la même entreprise, les modèles Mixtral7x8B, Mistral7B Instruct et Mistral7B, ont également une corrélation significative, ce qui peut être attribué à la similarité intrinsèque de leur architecture et de leur processus d'entraînement et sans doute aussi de leurs corpus de pré-entraînement. Parallèlement, les modèles Pythia 7B et Pythia 13B, qui ont été entraînés sur la même base de données, The Pile, se sont révélés être relativement corrélés, avec un coefficient de corrélation de 0,32. Ces résultats reflètent que les modèles issus de la même société, même avec des variations du nombre de paramètres, tendent à se corréler dans l'évaluation de notre tâche.

Ce qui est surprenant, c'est que par rapport à GPT-4, ChatGPT (gpt-3.5-turbo) est plus étroitement corrélé avec la série des modèles Mistral AI. En particulier, Mistral7B Instruct est même plus corrélé avec ChatGPT qu'avec Mistral7B. Cette observation pourrait nous amener à supposer que Mistral7B Instruct a été fine-tuné à partir de données de dialogue similaires à celles utilisées par ChatGPT, ou que les données destinées au fine-tuning ont été constituées partiellement à partir de dialogues de ChatGPT. Nous avons donc utilisé dans cette sous-section le coefficient de corrélation de Spearman pour examiner s'il existait une corrélation entre les scores des différents modèles, comme l'illustre la figure 2.8. Il est d'abord à noter que les modèles GPT-4 et ChatGPT présentent une corrélation relativement élevée de 0,707, ce qui implique qu'ils ont tendance à varier de manière similaire (l'isochronisme). De plus, développés par la même entreprise, les modèles Mixtral7x8B, Mistral7B Instruct et Mistral7B, ont également une corrélation significative, ce qui peut être attribué à la similarité intrinsèque de leur architecture et de leur processus d'entraînement et sans doute aussi de leurs corpus de pré-entraînement. Parallèlement, les modèles Pythia 7B et Pythia 13B, qui ont été entraînés sur la même base de données, The Pile, se sont révélés être relativement corrélés, avec un coefficient de corrélation de 0,32. Ces résultats reflètent que les modèles issus de la même entité, même avec

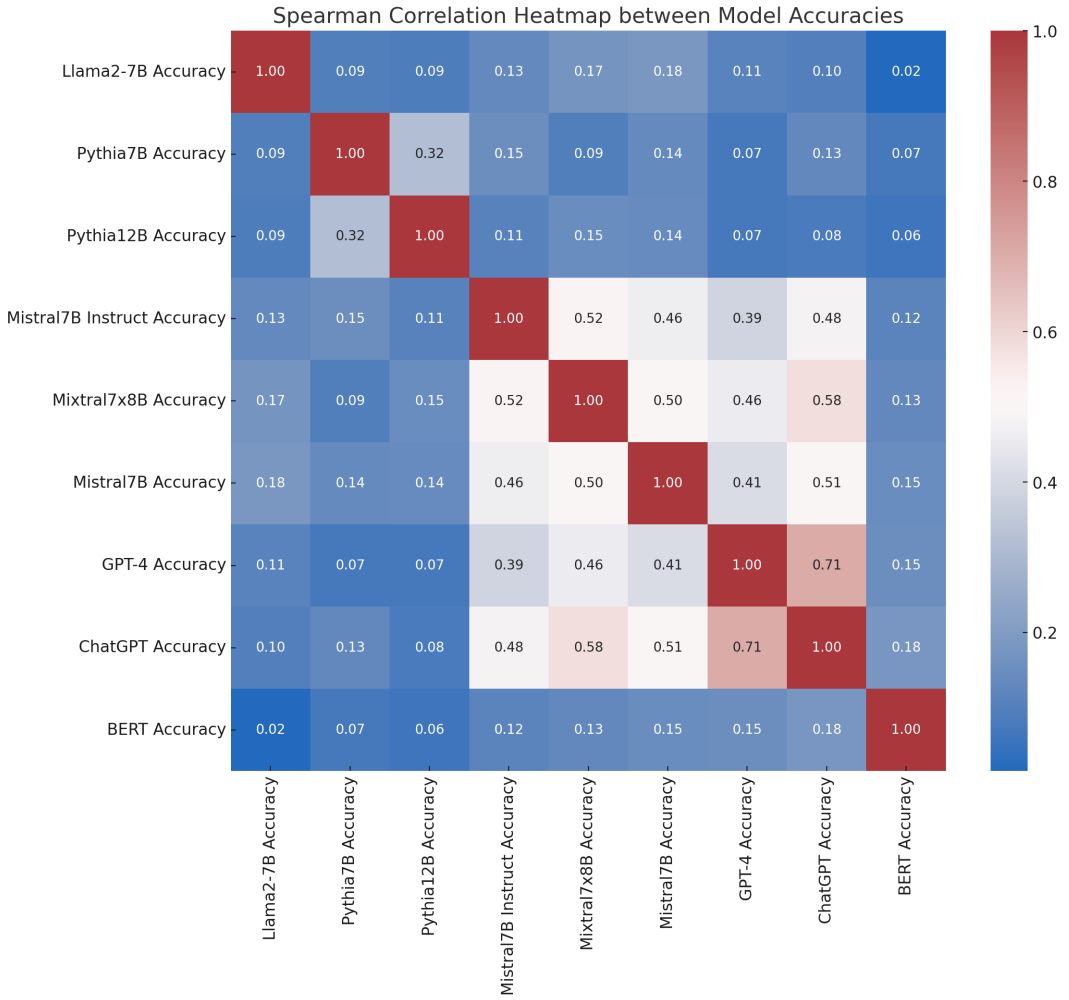


FIGURE 2.8 – Matrice de corrélation entre les précisions de tous les modèles

des variations du nombre de paramètres, tendent à se corréler dans l'évaluation de notre tâche.

Ce qui est surprenant, c'est que par rapport à GPT-4, ChatGPT (gpt-3.5-turbo) est plus étroitement corrélé avec la série des modèles Mistral AI. En particulier, Mistral7B Instruct est même plus corrélé avec ChatGPT qu'avec Mistral7B. Cette observation pourrait nous amener à supposer que Mistral7B Instruct a été fine-tuné à partir de données de dialogue similaires à celles utilisées par ChatGPT, ou que les données destinées au fine-tuning ont été constituées partiellement à partir de dialogues de ChatGPT.

2.4 Conclusion

Dans ce chapitre, nous avons répliqué les expériences de Chang et al. (2023) en utilisant plusieurs modèles ouverts pour prédire un nom propre manquant dans un extrait littéraire. Notre *prompt* visait à guider les modèles vers une prédiction précise, en évitant les noms fréquemment prédits et en précisant les instructions pour minimiser les réponses superflues.

Nos résultats, mesurés également en termes de précision de complétion de nom,

étaient globalement inférieurs à ceux de Chang et al. (2023). C'est-à-dire que les modèles d'OpenAI, et en particulier GPT-4, restent encore les plus performants pour la tâche.

Les analyses de Chang et al. (2023) ont été généralement répliquées par nos expériences. Celles-ci ont porté sur les facteurs susceptibles d'influencer la précision des prédictions, tels que le statut des droits d'auteur, le genre littéraire et la popularité des livres. L'analyse des scores moyens pour les livres soumis et non soumis à droits d'auteur a démontré que les œuvres dans le domaine privé sont moins représentés dans les prédictions correctes des modèles. De plus, les livres de science-fiction et de *fantasy* sont plus facilement mémorisés par les modèles, probablement en raison de leur utilisation fréquente de termes spécifiques. En observant la corrélation majoritairement positive entre les performances des modèles et la fréquence d'apparition des textes sur le web, on peut conclure que la présence des livres en ligne est un facteur déterminant. Enfin, le calcul des corrélations entre les modèles a révélé que ceux issus de la même entreprise présentent une corrélation dans le cadre de notre tâche.

En somme, nous avons eu des résultats semblables tout en répliquant les expériences de Chang et al. (2023) dans les modèles ouverts et ainsi confirmé leurs observations à partir des modèles fermés. Ce chapitre a également fourni un regard sur plusieurs facteurs qui pourraient entrer en jeu dans la performance des LLMs pour la tâche de complétion de noms, notamment le statut du copyright, le genre des livres et la popularité sur Internet.

CHAPITRE 3

VÉRIFICATIONS EMPIRIQUES DES EXPÉRIENCES DE CHANG ET AL. (2023)

Puisque, pour Chang et al. (2023), la capacité de mémorisation des livres a été évaluée uniquement par une seule métrique — *l'accuracy* de la prédiction du nom manquant — nous évaluerons dans ce chapitre en plus de profondeur la capacité de mémorisation précise des modèles pour les livres. Ces expériences servent à approfondir et à valider les résultats des expériences du chapitre précédent en fonctions de deux étapes principales : l'évaluation approfondie de la mémorisation et l'application en aval de la mémorisation.

3.1 Évaluation approfondie de la mémorisation

Cette section présente les expériences sur la trajectoire de mémorisation pendant l'entraînement des modèles et mesure directement comment les modèles apprennent et réutilisent les informations mémorisées.

3.1.1 *Checkpoints* d'OLMo et trajectoire de mémorisation

Objectif

Dans le chapitre précédent, nous avons vu que le nombre de *hits* pour les extraits des livres joue un rôle dans la réussite des modèles à prédire le nom manquant. Cependant, nous n'avons pas encore pu répondre à la question si pour mémoriser un livre, avoir accès au texte intégral est nécessaire ou si la mémorisation peut aussi avoir lieu à travers des extraits présents sur des sites web. Dans cette sections, nous présentons donc une nouvelle série d'expérience, pour étudier le taux de réussite à travers l'entraînement pour des livres du domaine public (pour lesquels on peut s'attendre à une présence du texte intégral) et sous droits d'auteur en fonction de leur popularité mesuré par le nombre de *hits*. Motivés par une série d'expériences qui ont examiné le comportement de la mémorisation lors de la phase de pré-entraînement (Biderman et al., 2023, 2024) et sur le conseil du chercheur Noah Smith [1] (co-directeur du projet OLMo) lors d'une échange en personne après un séminaire, nous avons décidé tester le modèle OLMo à différents stades de son entraînement, via des points d'arrêt (*checkpoints*). Cette démarche permet d'analyser l'évolution de la performance de prédiction du modèle au fur et à mesure de son entraînement en fonction du statut public/privé des livres et leur popularité sur le web.

1. https://nasmith.github.io

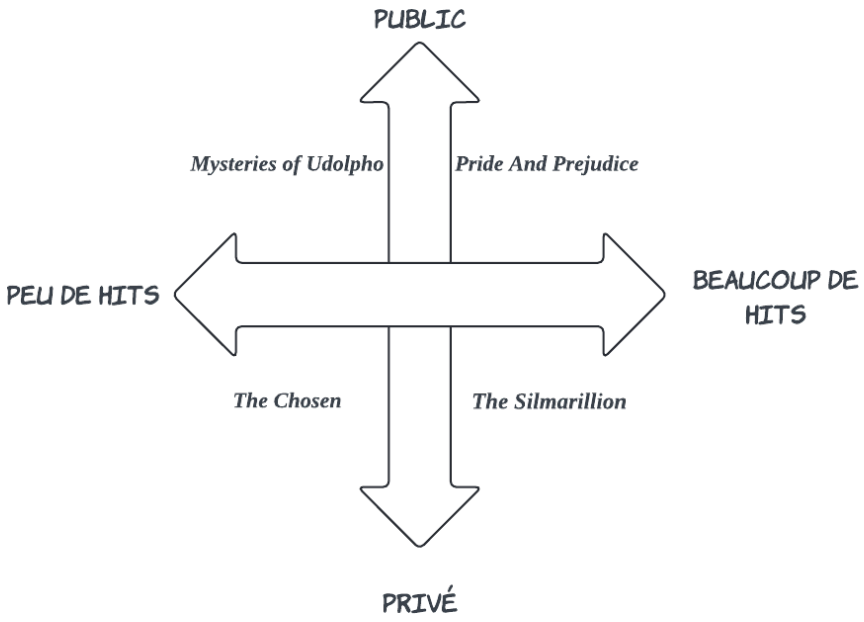


FIGURE 3.1 – Quatre Dimension

Méthode

Le modèle OLMo (Groeneveld et al., 2024) est le seul LLM qui a été entraîné sur des données entièrement publiques, le corpus Dolma (Soldaini et al., 2024). AllenAI a également publié de nombreux points d'arrêt (un point par 1000 étapes d'entraînement), qui permettent d'inspecter des états du modèle pendant l'entraînement.

Pour cette expérience il n'est intéressant d'effectuer des expériences pour l'ensemble des 571 livres en l'occurrence, sinon la tâche serait trop fastidieuse et consommatrice de temps, notamment lorsqu'il s'agit de multiplier les versions de modèles pour la prédiction d'un même livre : il faudrait télécharger autant de modèles OLMo qu'il y a des *checkpoints*, à savoir plus de 500 et donc l'expérience prendrait un temps d'exécution 500 fois supérieur à l'expérience initial avec ce modèle, c'est-à dire environ 500 jours (voir la section 2.2.1). C'est pour cela que, dans notre étude, nous nous sommes concentrés sur quatorze *checkpoints* — choisi à intervalle régulier — et quatre ouvrages particulièrement représentatifs, sélectionnés en fonction de deux dimensions, comme l'illustre la Figure 3.1 : le statut de droit d'auteur (public ou privé), le niveau de popularité (peu de hits ou beaucoup de hits). Ces ouvrages sont respectivement *the Mysteries of Udolpho*, *Pride and Prejudice*, *The Chosen* et *The Silmarillion*.

Nous avons décidé d'analyser aussi *Alice's Adventures in Wonderland* parce que ce livre est de loin le mieux mémorisé. Cependant, nous ne le prenons pas comme 'exemple typique' d'un livre du domaine public avec beaucoup de *hits* car en examinant les items expérimentaux de Chang et al. (2023) de plus près, ces items contiennent beaucoup d'indices qui permettent de reconnaître le livre, notamment des animaux qui sont des personnages mais qui n'ont pas été détectés en tant que

tels (voir l'exemple (1)). De plus, le livre contient un vocabulaire très typique. Ce même problème a déjà été souligné pour *Harry Potter and the Sorcerer's Stone* dans la section 2.3.4 ; et c'est d'ailleurs pour cette raison que ce livre a été exclu des analyses dans cette section. Il faut noter que l'œuvre *The Silmarillion* souffre aussi dans un certain degré de ce problème, mais quand-même beaucoup moins et n'ayant pas d'autre livre moins 'contaminé' dans la catégorie 'beaucoup de hits & privé' nous l'avons choisi tout de même pour cette expérience.

(1) a. 'Can't remember WHAT things?' said **the Caterpillar**. 'Well, I've tried to say "HOW DOTH THE LITTLE BUSY BEE," but it all came different!' [MASK] replied in a very melancholy voice.
b. It quite makes my forehead ache!' [MASK] watched **the White Rabbit** as he fumbled over the list, feeling very curious to see what the next witness would be like, '–for they haven't got much evidence YET,' she said to herself.
c. Oh, you can't help that,' said **the Cat** : 'we're all mad here. I'm mad. You're mad.' 'How do you know I'm mad?' said [MASK].

Résultat

Le résultat est alors présenté dans la figure 3.2. En effet, pour les ouvrages relevant du domaine public, tels que *the Mysteries of Udolpho*, *Pride and Prejudice* et *Alice's Adventures in Wonderland*, on observe une augmentation notable des scores de prédiction vers la fin de l'entraînement, notamment entre les steps 450 000 et 557 000. Cette augmentation apparaît comme une augmentation considérable dans la graphique. À partir de cette observation, il serait raisonnable d'avancer qu'à ce stade de l'entraînement, le modèle sont probablement entraînés avec des textes intégraux d'œuvres libres, comme ceux disponibles dans les projets les plus réputés tels que Project Gutenberg ou LitBank, etc. En contraste, pour les deux autres ouvrages protégés par le droit d'auteur, *The Chosen* et *The Silmarillion*, leur performance a évolué de manière continue et stable tout au long de la période d'entraînement, sans présenter une augmentation aussi marquée et soudaine, qui serait habituellement représentée par une courbe très abrupte. Par exemple, dès le début de la phase de pré-entraînement, à partir du step 50 000 (presqu'au tout début du pré-entraînement), le modèle OLMo a réussi à prédire un nom propre masqué dans les items de *The Silmarillion*. Leurs précisions ont légèrement fluctué mais s'est maintenue relativement stable tout au long de la phase d'entraînement, jusqu'à la fin, bien qu'il y ait eu quelques bonnes prédictions supplémentaires. Cela pourrait être soutenu par l'hypothèse que des extraits ou des citations de ce livre, sont disséminés dans divers sous-corpus et répartis tout au long de la phase de pré-entraînement. De plus, il est évident que l'influence de la popularité sur le web, mesurée par le nombre de 'hits', joue aussi un rôle important de l'évolution surtout pour les ouvrages sous droit d'auteur. Cela est particulièrement observable pour l'œuvre *The Silmarillion*, dont la popularité croissante sur le web est associée à des fluctuations de scores prédictifs plus prononcées (vu que la courbe de *The Silmarillion* s'avère plus fluctué par rapport à celle de de *The Chosen*).

La recherche plus poussée dans cette sous-section serait de vérifier si le score de prédiction qui reste inchangé dans certains steps représente tout le temps la même prédiction pendant l'entraînement. Envisageons un scénario où, pour plusieurs steps, le score est le même, mais les éléments bien prédits ne sont pas toujours les mêmes.

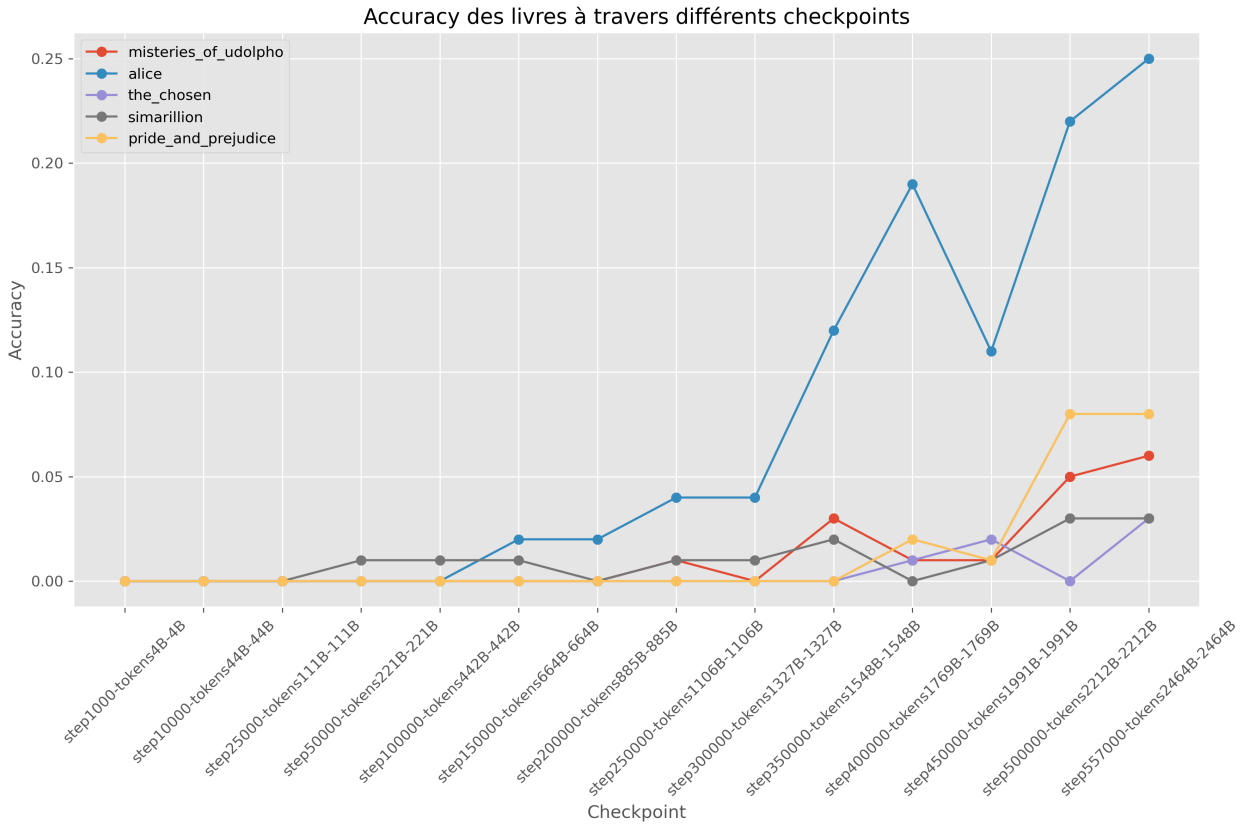


FIGURE 3.2 – Évolution des scores à travers divers checkpoints

Analyse et perspective futur

Nous avons donc généré des tableaux détaillés qui répertorient l'ensemble des items correctement prédits par les différents checkpoints du modèle. Ces tableaux sont présentés en annexe, classés par livres sélectionnés. Comme prévu, nous avons observé que les modèles ne sont pas stables tout au long de la phase de pré-entraînement. Ils mémorisent et oublient simultanément des données, ce qui se manifeste par des modifications significatives des items bien prédits.

Afin d'évaluer la stabilité de la mémorisation des modèles au cours de l'entraînement, nous avons proposé un algorithme de calcul du ***score d'oubli moyen***. L'algorithme vise à mesurer l'indice d'oubli du modèle pour chaque OriginalIndex (identifiant d'item du livre), en supposant que chaque OriginalIndex, dès qu'il apparaît pour la première fois à un checkpoint, devrait ré-apparaître dans tous les checkpoints suivants. D'après nous, l'absence d'un OriginalIndex dans un checkpoint ultérieur indique que le modèle a oublié cet index. Normalement, chaque index correspond à un extrait avec un nom manquant sur cent items pour chaque livre.

Nous vous expliquons l'implémentation de cet algorithme : il faut d'abord trouver pour chaque OriginalIndex i, le checkpoint C_i où i est observé pour la première fois. Ensuite, on va détecter l'oubli de cet indice : pour chaque OriginalIndex i et pour chaque checkpoint $c \in C$ après C_i, on vérifie si i est manquant. Par exmple, si i est manquant dans un checkpoint c après C_i, on va augmenter N_i de 1. Ici, N_i signifie le nombre de checkpoints après C_i où l'OriginalIndex i est manquant. Tout cela permet

de calculer dans la première partie le taux d'oubli T_i pour chaque OriginalIndex i :

$$T_i = \frac{N_i}{N - \text{position de } C_i}$$

où N est le nombre total de checkpoints et "position de C_i" fait référence à l'index de C_i dans l'ensemble C (par exemple, si C_i est le troisième checkpoint, sa position est 3).

Le ***score d'oubli moyen*** du modèle est finalement obtenu en calculant la moyenne des taux d'oubli de tous les OriginalIndexes i :

$$Score\_oubli_{\text{moyen}} = \frac{\sum_{i=1}^{n} T_i}{n}$$

où n est le nombre total d'OriginalIndexes considérés.

Les résultats sont présentés dans le Tableau 3.1. Une analyse rapide du tableau révèle que les modèles ont tendance à oublier les informations provenant de livres du domaine privé. Cela est démontré par les scores élevés de *The Chosen* et *The Silmarillion*, qui sont supérieurs à ceux de *Pride and Prejudice* et *The Mysteries of Udolpho*. Cette observation est pertinente, car les modèles sont généralement entraînés sur des livres dans le domaine publique, de la manière dont des extraits et des résumés sont largement répandus sur Internet et ainsi distribués partout pendant l'entraînement, plutôt que des livres entiers.

Livres Représentatifs	**Score d'oubli moyen**
pride and prejudice	0.205
the silmarillion	0.458
the chosen	0.292
the mysteries of udolpho	0.190

TABLE 3.1 – Comparaison de score d'oubli moyen pour les livres

Or, ces résultats ne sont pas tellement convaincant car il paraît un peu pénible d'établir une ligne de base de l'oubli pour les modèles au fil des checkpoints, ce qui permettrait de réaliser des comparaisons plus fines. Nous réfléchissons un peu à une approche inspirée de la recherche de Tirumala et al. (2022b), qui consiste à prendre un checkpoint du modèle, à y intégrer un batch des données non-disponibles dans le corpus d'entraînement afin que le modèle puisse s'entraîner dessus, puis à reprendre l'entraînement standard sur le jeu de données d'entraînement. On évalue ensuite comment la mémorisation se dégrade pour ce batch spécial. L'idée principale serait d'injecter des données non disponibles dans le corpus d'entraînement à différents checkpoints et d'observer comment la mémorisation évolue pour ces données.

De toute façon, cette question reste intéressante à explorer dans les futurs travaux. Compte tenu des contraintes de temps et de ressources, nous devons nous arrêter à ce stade sans approfondir davantage ces aspects.

3.1.2 Évaluation des préférences des personnages

Dans cette sous-section, nous examinons la performance des modèles dans le cadre de l'identification des personnages principaux et secondaires. Plus précisément, nous nous interrogeons sur la capacité des modèles à prédire plus efficacement les noms des personnages principaux ou secondaires. Cette question est plutôt posée en

comparant les modèles ouverts et fermés. En vue de détecter les noms des personnages de chaque livre, nous identifions les noms masqués les plus associés aux livres spécifiques en calculant l'information mutuelle positive (PPMI) entre chaque paire nom de personnage et livre.

Extraction des personnages principaux et calcul de PPMI

Nous avons traité directement les jeux d'items, pour 571 livres, chacun représenté comme cent extraits de chaque livre. Nous avons ensuite pré-traité tous les extraits en utilisant l'annotation pos-tagging et NER de la version anglaise de *spaCy* pour identifier et extraire les noms propres qui correspondent probablement aux noms de personnages. Compte tenu de la facilité du calcul, la structure finale des données constitue en une dictionnaire `book_data` où chaque clé est le titre du livre et chaque valeur est une liste de mots (noms de personnage) qui apparaît dans les extraits du livre, voir l'exemple (2).

(2) Un exemple de dictionnaire `book_data` montrant l'association entre les titres des livres et leurs personnages :

{
'evanovich_one_for_the_money_One_for_the_Money' : ['Rex', 'Vinnie', 'Morelli', 'Carmen', 'Morelli',...],
'faulkner_a_fable_A_Fable' : ['Byron', 'Marthe', 'Bridesman', 'Caesar', 'Harry',...],
'metalious_peyton_place_Peyton_Place' : ['Constance', 'Rodney', 'Ted', 'Allison', 'Kenny',...],
'king_tommyknockers_The_Tommyknockers' : ['Gard', 'Nancy', 'David', 'Bobbi', 'Moss'],
'432_the_ambassadors' : ['Jeanne', 'Waymarsh', 'Chad', 'Jeanne', 'Chad',...]
}

Nous avons donc pu calculer la fréquence de chaque personnage pour chaque livre et à travers l'ensemble du corpus. Soit C l'ensemble de tous les personnages présents dans au moins un livre. Pour chaque livre $b \in \mathcal{B}$, nous définissons :

— $f_{c,b}$: Fréquence du personnage c dans le jeu d'items du livre b
— f_c : Fréquence totale du personnage c à travers les jeux d'items de tous les livres
— N : Nombre total des personnages à travers les jeux d'items de tous les livres
— N_b : Nombre total des personnages dans l'item du livre b

Le score de PPMI pour un personnage c dans un livre b est calculé à l'aide de la formule suivante :

$$\text{PPMI}(c, b) = \max\left(\log_2\left(\frac{P(c \mid b)}{P(c)}\right), 0\right)$$

où

$$P(c \mid b) = \frac{f_{c,b}}{N_b} \quad \text{et} \quad P(c) = \frac{f_c}{N}$$

Pour chaque livre b, le personnage principal est identifié comme le personnage c^* qui maximise le score PPMI :

$$c^* = \arg\max_{c \in C} \text{PPMI}(c, b)$$

cette approche nous permet de déterminer le personnage le plus saillant dans la structure narrative du livre en fonction de sa présence textuelle et de son particularité. Après avoir testé notre méthode sur plusieurs œuvres littéraires classiques, nous avons observé que les personnages principaux identifiés correspondent bien aux protagonistes connus des histoires. Par exemple, dans *Alice's Adventures in Wonderland*, le personnage 'Alice' est identifié avec le score PPMI le plus élevé.

Deux indicateurs d'évaluation de la performance

Après avoir identifié le les noms de personnages principaux pour chaque livre, nous avons alors conçu deux approches de calcul pour mesurer comment les modèles réussissent à prédire les personnages principaux et secondaires, comparés aux prédictions globales et attendues :

Métrique I : Représentativité des Personnages Principales parmi les Prédictions Réussies $P_{\text{main\_tout}}$ Il s'agit du pourcentage de noms de personnages principaux parmi les noms correctement prédits, qui permet de voir si les prédictions réussies sont dominées par les personnages principaux. :

$$P_{\text{main\_tout}} = \frac{\text{Nombre de noms de personnages principaux correctement prédits}}{\text{Nombre total de noms correctement prédits}}$$

Métrique II : Efficacité Globale de la Prédiction des Personnages Principaux ou Secondaires($P_{\text{main/non-main}}$) Cet indicateur évalue la performance du modèle pour prédire les noms de personnages principaux ou secondaires par rapport au nombre total de noms de personnages principaux ou secondaires attendus(masqués), pour chaque livre :

$$P_{\text{main/non\_main}} = \frac{\text{Nombre de personnages principaux ou secondaires bien prédits}}{\text{Nombre total de personnages principaux ou secondaires attendus}}$$

En générale, la première méthode permet de quantifier la représentativité des personnages principaux parmi les prédictions réussies, tandis que la seconde se concentre sur l'efficacité globale du modèle à prédire les personnages principaux, sans considérer leur fréquence brut dans les données.

Tout en appliquant les deux méthodes de calcul décrites précédemment, nous avons pu respectivement obtenir deux scores pour chaque livre par modèle. Pour évaluer de manière plus globale la performance du modèle, nous avons décidé de calculer la moyenne des deux types de scores séparément. Toutefois, afin d'assurer une évaluation significative, cette moyenne a été calculée en se limitant aux livres pour lesquels le modèle a prédit avec succès au moins deux noms masqués, soit les livres ayant un score d'*accuracy* total supérieur à 1% (0.01). Ce seuil permet de garantir que les évaluations sont basées sur des performances minimales acceptables.

Les scores moyens pour chaque indicateur sont calculés comme suit :

$$S_{\text{random}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} P_{\text{random},b}$$

où S_{random} peut représenter le score moyen pour l'un des indicateurs ($S_{\text{main\_tout}}$, S_{main}, ou $S_{\text{non-main}}$). De plus, nous avons défini $\mathcal{B}$ comme l'ensemble des livres pour lesquels le modèle a prédit au moins deux noms masqués ou bien où le score d'*accuracy* total est supérieur à 0.01. $P_{\text{random},b}$ est le score de l'indicateur spécifique pour le livre b.

Résultat et Analyse

Model	$S_{\textbf{main\_tout}}$	$S_{\textbf{main}}$	$S_{\textbf{non-main}}$
Chatgpt	0.356402	0.091376	0.047302
GPT4	0.321543	**0.155466**	**0.094823**
Mistral7B	0.397945	0.065205	0.030548
Mistral7B Inst	0.309531	0.051094	0.030156
Mixtral7x8B	0.342766	0.047234	0.030532
OLMo7B	**0.509615**	0.043462	0.015000

TABLE 3.2 – Résultat d'analyse des personnages

Dans le tableau 3.2, on pourrait observer que GPT-4 demeure le meilleur modèle pour la prédiction des personnages, qu'ils soient principaux ou secondaires, comme le suggèrent les scores $S_{\text{non-main}}$ et S_{main}. Ces résultats sont en effet en accord avec les scores de performance calculés dans le chapitre précédent. Aussi, pour tous les modèles, il existe une tendance à prédire correctement le personnage principal par rapport au personnage secondaire.

Toutefois, pour le résultat $S_{\text{main\_tout}}$, c'est le modèle OLMo7B qui obtient le score le plus élevé. Cela indiqu'une forte tendance à inclure les personnages principaux dans ses prédictions réussies, vu presque la moitié de tous les personnages bien identifiés. On aimerait alors supposer que le modèle a potentiellement une tendance à privilégier les personnages principaux lors de la prédiction, même si cette stratégie ne peut pas toujours aboutir à des résultats corrects. Cela soulève donc la question de la tendance du modèle à 'deviner' les noms de personnages principaux, plutôt que de les déduire de manière plus analytique. Afin d'approfondir d'étudier cette hypothèse, il serait intéressant dans le futur de procéder à une analyse des erreurs commises par le modèle, en examinant les cas où des noms de personnages ont été mal prédits. En outre, il serait également pertinent d'examiner si cette prédilection pour les personnages principaux est partagée par d'autres LLMs sauf OLMo7B.

3.1.3 Impact des répétitions au niveau des extraits

Des études, comme celles de Carlini et al. (2023), ont montré qu'une forte présence de contenus dupliqués dans le corpus d'entraînement est corrélée à la mémorisation. Nous avons effectué des expériences dans cette sous-section pour évaluer si le modèle a tendance à prédire les extraits qui se répètent dans le corpus d'entraînement.

Méthode

Nous avons évalué cette problématique en utilisant le corpus complètement libre - Dolma, avec lequel le modèle OLMo est entraîné. Pour chaque livre, nous avons procédé à la sélection aléatoire de 10-grammes à partir de chaque extrait textuel. Nous avons ensuite effectué des requêtes sur l'outil *infini-gram* (Liu et al., 2024), un moteur capable de traiter les requêtes n-gram/∞-gram sur des corpus de texte massifs, y compris Dolma. L'interface Web de *infini-gram* est accessible via le site [2]. Nous avons utilisé son point de terminaison API [3], qui permet de servir les requêtes n-gram pour trouver le nombre d'occurrences correspondant exactement à la même

2. https://huggingface.co/spaces/liujch1998/infini-gram
3. https://infini-gram.io/api_doc

séquence de caractères. Selon la documentation, il s'agit d'extraire l'argument 'count' afin de déterminer le nombre de répétitions.

Résultat

Après avoir collecté le nombre d'occurrences de chaque extrait, plus précisément les 10-grammes, dans le corpus Dolma, nous avons donc pu procéder à l'analyse des données. La Figure 3.3 est la boxplot qui montre la distribution des valeurs de dolma_search pour chaque niveau de prédiction. Les deux boîtes représentent la distribution de dolma_search pour les prédictions correctes et incorrectes. Cependant, on ne peut pas dire à partir de cette figure que la fréquence d'occurrence ne présente pas de corrélation significative avec la précision des prédictions, c'est-à-dire entre le nombre d'occurrences d'un dix-gramme (dolma_search) et le résultat prédiction égale à True. Même pour les occurrences supérieures à 2 500 fois, le modèle n'est pas en mesure de correctement prédire cet élément. Pour essayer d'en tirer une conclusion plutôt ferme, nous avons donc analyser des valeurs extrêmes (outliers) pour les 'dolma_search' dans les prédictions 0 et 1. Les tableaux suivants comprennent les 5 plus grandes valeurs de 'dolma_search' pour chaque catégorie de prédiction, accompagnées de leur FileID, predit, gold, et dix_grammes.

FileID	Predit	Gold	Dolma Search	Dix Grammes
113_the_secret_garden	The voice	Dickon	2713	as if it was the most natural thing in the
11_alices_adventures_in_wonderland	""	Alice	503	went straight on like a tunnel for some way, and
2814_dubliners	""	EVELINE	481	the darkness I saw myself as a creature driven...
2489_moby_dick	Cruise	Cato	479	cherish very nearly the same feelings towards ...
lahaye_desecration_Desecration	John	Rayford	477	once was lost but now am found, was blind but

TABLE 3.3 – Top 5 Dolma Search pour la Prédiction 0

Dans le tableau 3.3, nous constatons que les dix grammes dont l'occurrence est la plus élevée est la phrase "as if it was the most natural thing in the" qui renvoie à une construction très courante dans le langage quotidien.

FileID	Predit	Gold	Dolma Search	Dix Grammes
11_alices_adventures_in_wonderland	Alice	Alice	276	so many out-of-the-way things had happened lately,
11_alices_adventures_in_wonderland	Alice	Alice	138	heard the Rabbit just under the window, she su...
11_alices_adventures_in_wonderland	Alice	Alice	107	all stopped and looked at her, and the Queen said
11_alices_adventures_in_wonderland	Alice	Alice	104	together, Alice heard the King say in a low vo...
11_alices_adventures_in_wonderland	Alice	Alice	95	looking for eggs, as it happens ; and if I was,

TABLE 3.4 – Top 5 Dolma Search pour la Prédiction 1

Dans le tableau 3.4, nous observons que tous les enregistrements proviennent du livre *11_alices_adventures_in_wonderland* et concernent la prédiction "Alice". Les dix-grammes qui reviennent fréquemment dans un contexte narratif spécifique, comme celui de *Alice's Adventures in Wonderland*, présentent des scores élevés de dolma_search, surtout lorsqu'elles impliquent des personnages ou des éléments emblématiques. Cela a de nouveau prouvé que la spécificité textuelle joue un rôle dans la bonne prédiction des LLMs.

Cependant, l'analyse des valeurs aberrantes pour les prédictions incorrectes met en lumière une limitation de notre expérience. Les phrases courantes et populaires peuvent avoir des scores élevés de dolma_search, même si elles ne sont pas pertinentes pour la classification spécifique que nous cherchons à établir. En cause, dix-grammes comme déclencheurs de recherche peut être biaisé vers des expressions couramment utilisées. Par conséquent, bien que le dolma_search soit un bon indicateur

de la fréquence ou de la popularité des extraits littéraires dans le corpus, il faut l'interpréter avec prudence. Pour rendre l'expérience plus nuancée, nous avons ensuite effectué un test de Chi-2 d'indépendance. Les résultats du test sont présentés dans le Tableau3.5. Globalement, il y a une association statistiquement significative entre la valeur de dolma_search supérieure à 0 et la prédiction positive, puisque le p-valeur, d'environ 4.55×10^{-12}, est considérablement inférieur au niveau de significativité habituel (par exemple, 0.05). Cependant, la table de contingence révèle le même constat que malgré une valeur de dolma_search supérieure à 0, le nombre de prédictions positives ne l'emporte pas.

	Prediction	
	False	True
dolma_search = 0	40784	71
dolma_search > 0	16162	83

TABLE 3.5 – Résultats du test de Chi-2 d'indépendance

Nous avons donc pensé à appliquer la méthode de machine-learning, telles que la régression logistique, pour quantifier l'impact de deux variables clés. Cependant, notre jeu de données est fortement déséquilibré, avec beaucoup plus de prédictions incorrectes que correctes. Pour atténuer ce déséquilibre, la technique SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002) pourrait être appliquée. SMOTE génère des exemples synthétiques pour équilibrer les classes minoritaires. Toutefois, vu l'ampleur tellement profonde du déséquilibre, cette méthode pourrait ne pas suffire.

Une alternative consiste à entraîner le modèle de manière itérative, en utilisant un échantillonnage aléatoire des prédictions incorrectes, similaire à l'algorithme de boosting. Faute de temps, ces méthodes seront explorées dans les travaux futurs.

3.1.4 Impact d'affinage du modèle

Cette sous-section finale est consacrée à l'expérience d'affinage du modèle Mistral7B afin d'évaluer l'impact de la méthode de l'affinage des instructions (*prompt-tuning*) sur sa performance de mémorisation des livres.

Méthode de l'affinage

Pour constituer les données d'affinage, nous avons sélectionné des textes issus de livres du domaine public, en supposant que le modèle avait déjà été préalablement entraîné sur la majorité de ces œuvres. Cela est important pour tester comment le modèle réagit à un raffinage sur des données qu'il a probablement déjà vues. De plus, les livres ont été classés en deux catégories basées sur leur score d'*accuracy* obtenu dans l'expérience du deuxième chapitre. Cette classification permet de différencier les livres bien mémorisés (score > 0,05) de ceux mal mémorisés (score nul). Nous avons alors sélectionné de manière équilibrée les livres de ces deux catégories.

Les données ont finalement été structurées au format JSON, où chaque entrée comprend :

— **input** : le texte à compléter.
— **output** : la réponse attendue du modèle.

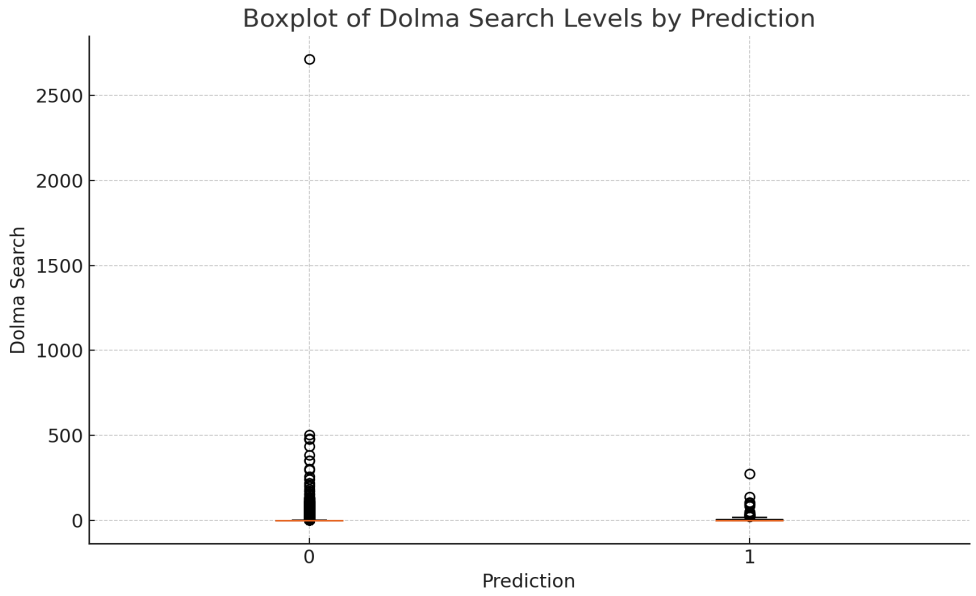


FIGURE 3.3 – Distribution des occurences du corpus Dolma par prédiction (1 = Prédiction correcte, 0 = Prédiction incorrecte)

— **instruction** : des directives spécifiques, similaires à celles utilisées dans l'expérience de Chang et al. (2023).

Un exemple de format d'entrée est illustré ci-dessous :

```
[
{
  "input": "You want breakfast, [MASK], or piss me off?",
  "output": "<name>Gard</name>",
  "instruction": "You have seen the following passage ..."
},
...]
```

En ce qui concerne la méthode d'affinage, nous avons notamment employé Lora (Hu et al., 2021), une technique de quantification de modèles qui est disponible sur la bibliothèque Python *peft* [4]. Le modèle ainsi affiné a été intégré et est accessible sur le site de mon compte de Hugging Face [5], où il est présenté avec les résultats de l'expérience d'affinage. Pendant l'affinage, nous avons recouru à la plate-forme wandb [6] pour inspecter l'expérience.

Pendant la configuration de l'affinage, nous avons utilisé l'API `transformers.Trainer`. Voici quelques paramètres clés ci-dessous :

— **num_train_epochs** : fixé à 1 pour une seule époque d'affinage.
— **per_device_train_batch_size** : 4, pour gérer efficacement la mémoire pendant l'entraînement.
— **learning_rate** : 2e-4, pour un ajustement approprié des paramètres du modèle.

4. https://pypi.org/project/peft/
5. https://huggingface.co/LivevreXH/mistral_finetuned_items_livres/tree/main
6. https://wandb.ai/site

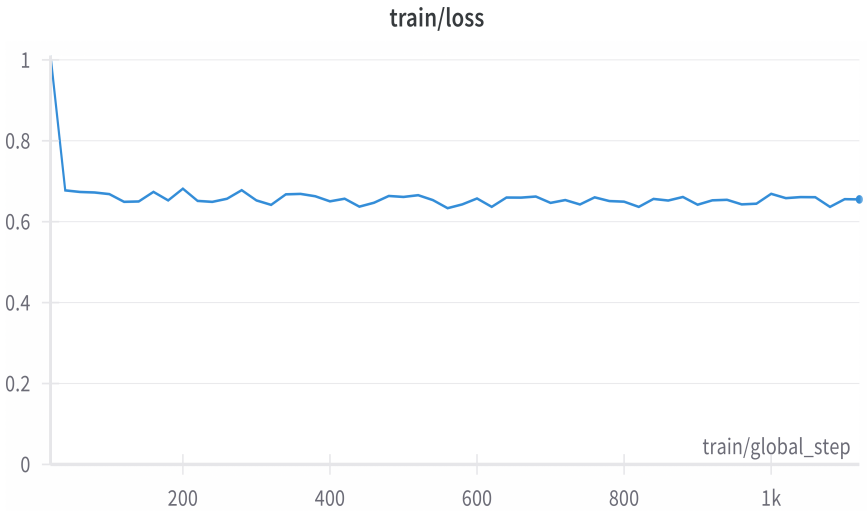


FIGURE 3.4 – Évolution de la perte pendant l'affinage

— **fp16** : activé pour réduire la consommation de mémoire et accélérer l'entraînement.
— **max_steps** : Limité à -1, ce qui signifie que l'entraînement continue jusqu'à la fin des époques spécifiées.
— **optim** : L'optimiseur utilisé est "paged_adamw_8bit", adapté aux grands modèles pour améliorer la performance et l'efficacité mémoire.
— **report_to** : "wandb" pour la visualisation et l'analyse des résultats en temps réel.

Résultats et limites

L'évolution de la valeur de perte est illustrée à la Figure 3.4. Il est observable que cette dernière diminue de manière significative uniquement pendant les premières étapes. Cependant, pour les étapes restantes de l'affinage, la valeur de perte ne varie pas considérablement et se stabilise autour de 0,6. Nous avons soumis le modèle affiné à la même tâche pour l'ensemble de nos 571 livres, et nous n'avons pas constaté d'amélioration significative par rapport au modèle de base, Mistral 7B, en termes de score d'*accuracy*. Le score d'*accuracy* moyen du modèle de base Mistral 7B est de 0.0467, tandis que le modèle affiné atteint un score de 0.0512. On peut suggérer que l'affinage n'a pas apporté de gains substantiels en termes de mémorisation des passages des livres testés. Il serait judicieux d'effectuer une évaluation plus approfondie de ce modèle affiné ou d'explorer d'autres méthodes d'affinage. Cela inclurait de prendre en compte non seulement l'affinage par le prompt lui-même, mais également l'affinage par le livre dans son ensemble.

3.2 Application en aval de la mémorisation

Dans cette section, nous avons mené une expérience comme tâches en aval pour explorer l'application de la mémorisation. L'objectif de cette étude est d'appliquer la mémorisation des livres dans d'autres tâches finalisés. On aimerait vérifier si on peut constater la même tendance que le résultat de l'expérience de Chang et al. (2023).

I will give you a sentence with 10 words from a book that you have memorized in your training corpus. Please provide the next sequence of up to 20 words, enclosed in <sent> and </sent>. Make your best guess, even if uncertain.

Examples:

Input: "He shuts the box and slips it into his pocket"
Output: <sent>when a knock sounds on the door. It is the headwaiter himself who enters with the drinks.</sent>

Input: "She heard the crunch and crackle of a bag of"
Output: <sent>fried plantains being crushed in her pocket. Her other pocket contained a small rubber ball, some string, a sliver of</sent>

Now, please provide the output surrounded by <sent> and </sent> without any explanation for the following input:
Input:

FIGURE 3.5 – Template de prompt pour la génération des continuité des textes

L'hypothèse est que un LLM permet de retenir et puis reproduire les séquences textuelles des livres plus populaires.

3.2.1 Génération prédictive des séquences textuelles

Prompt

Dans le cadre de cette expérience, la première étape consistait à concevoir un prompt spécifique. Nous avons largement adopté le modèle de l'expérience précédente, qui consiste à fournir deux exemples avant de faire une prédiction. Pour délimiter la séquence continue fournie par le modèle, nous avons utilisé la balise <sent></sent>. De plus, nous avons fixé une limite à la longueur du texte prédit, en ne ne dépassant pas les 20 premiers mots. Concernant l'entrée de chaque itération, nous avons utilisé les premiers 10-grammes des éléments après la substitution du vrai nom par le token manquant. L'exemple de prompt est fourni dans la figure 3.5.

Méthode d'évaluation

Pour évaluer la mémorisation, nous comptons nous servir dans cette expériences du framework introduit par Carlini et al. (2021) fondé sur la k-extractability. Dans le cadre de notre expérience, nous avons légèrement adapté la définition classique de la k-extractability pour l'appliquer à notre jeu d'évaluation. Au lieu de vérifier si une séquence spécifique (S) est présente dans le corpus d'entraînement du modèle, nous nous concentrons sur des extraits de textes réels tirés des livres de notre jeu d'évaluation.

Une séquence de caractères (S) est considérée comme k-extractible dans notre contexte si elle satisfait les conditions suivantes :

1. **Existence dans le livre de l'ensemble d'évaluation :** La séquence S doit être un extrait authentique du livre que nous utilisons pour tester la mémorisation.
2. **Reproduction par le modèle :** Le modèle doit générer la séquence S en réponse à un prompt composé des k tokens précédents.

Par exemple, considérons l'extrait du livre *Alice's Adventures in Wonderland* : 'So they got their tails fast in their mouths'. Si en fournissant le prompt 'So they got' (qui constitue les 3 premiers tokens), le modèle génère exactement 'their tails fast in their mouths', alors cette séquence est dite 6-extractible. Cela signifie que le modèle a mémorisé et reproduit fidèlement cet extrait, même si nous ne vérifions pas explicitement s'il est présent dans le corpus d'entraînement du modèle.

Pour quantifier la fidélité de la mémorisation, la formule du score de mémorisation est inventée comme ceci :

$$\text{score}(M, N) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(S_{M+i} = G_{M+i})$$

— M représente la longueur du prompt en nombre de tokens.
— N est la longueur de la séquence générée par le modèle que nous comparons.
— S_{M+i} est le i-ème token de la séquence attendue (l'extrait réel du livre) après le prompt.
— G_{M+i} est le i-ème token généré par le modèle après le prompt.
— $\mathbf{1}(x)$ est une fonction d'activation qui vaut 1 si l'assertion x est vraie et 0 sinon.

Résultat

L'expérience s'est effectué sur les quatre livres représentatifs ainsi que *Alice's Adventures in Wonderland*, pareilles à ceux utilisés dans l'expérience précédente. Les résultats sont présentés dans le Tableau 3.6. Dans ce tableau, le mémo-score représente la moyenne de cent essais par livre. Cependant, le mémo-score ne permet pas de déterminer avec certitude si le modèle a mémorisé un livre donné ou non. Il faut essayer de trouver la ligne de base. C'est pourquoi nous avons sollicité l'aide d'une de nos amies, une jeune romancière talentueuse, Jingyi. Elle nous a fourni un brouillon inédit de son prochain roman, rédigé en chinois. Ce manuscrit n'ayant pas encore été publié, nous supposons qu'il est inconnu des modèles de langue que nous testons. Nous avons traduit ce texte en anglais en utilisant l'outil de traduction DeepL[7]. À partir du manuscrit traduit, nous avons sélectionné 100 extraits aléatoires. Nous avons soumis ce manuscrit à la même tâche de prédiction. Malgré quelques répétitions aléatoires de ponctuations, le score de mémorisation demeure toujours très bas, voir le résultat dans le tableau 3.6.

En comparant les résultats obtenus avec la ligne de base, nous pouvons dire que certains extraits des livres représentatifs ont été correctement mémorisés par les modèles ChatGPT et GPT-4o. De plus, nous avons constaté que l'amplitude des mémo-scores est étroitement corrélée à celle des scores d'*accuracy* pour ces livres.

Néanmoins, cinq livres ne normalement suffissent pas à obtenir un résultat bien clair et significatif, il serait préférable d'effectuer cette tâche dérivée sur l'ensemble de 571 livres. Concernant la méthode d'évaluation, nous trouvons aussi pertinent de nous concentrer sur la fréquence des scores maximaux, c'est-à-dire de déterminer

7. https://www.deepl.com/fr/translator

Matériaux testés	**Mémo-score ChatGPT**	**Mémo-score GPT4o**
alices adventures in wonderland	0.357	0.328
pride and prejudice	0.196	0.098
the silmarillion	0.074	0.064
the chosen	0.047	0.040
the mysteries of udolpho	0.038	0.033
manuscrit romancier	0.005	0.008

TABLE 3.6 – Comparaison de score moyen de mémorisation pour livres représentatifs

combien de fois le modèle a tout réussi à prédire correctement la suite du prompt mot à mot, puisque nous avons observé plusieurs cas où le modèle a obtenu un score de 1.0 dans les fichiers du résultat. Par ailleurs, il serait également intéressant d'examiner l'impact de l'apparition des noms de personnages sur la prédiction des séquences continues. Cela pourrait être réalisé en calculant la corrélation entre la présence des noms de personnages et la performance de prédiction, ainsi qu'en explorant d'autres mesures pertinentes.

CHAPITRE 4

APPLICATION À DES MODÈLES POUR LE FRANÇAIS

4.1 Introduction

L'étude de Chang et al. (2023) portait sur les textes anglophones, mais la même méthode peut être appliquée au français, et c'est ce travail que nous allons présenter dans ce chapitre. Nous chercherons également à identifier les principales raisons de la différence potentielle avec le résultat de l'anglais, que ce soit liée au tokenizer des LLMs ou à d'autres facteurs.

4.2 Adaptation au français

Afin d'adapter les expériences au français, il serait d'abord préférable de sélectionner des Grands modèles de langue (LLMs) entraînés sur des corpus français de taille suffisamment importante, afin d'obtenir des résultats relativement comparables. Parallèlement, il serait nécessaire de générer nos propres items en français, c'est-à-dire des extraits littéraires en français qui seront introduits dans les prompts pour la tâche de prédiction par les modèles.

4.2.1 Modèles français

Étant donné que les expériences de réplication en anglais nous avons appris que plusieurs modèles ont des performances nettement inférieures, nous avons décidé de garder seulement les deux modèles ouverts, Olmo et Mistral 7B, dans les expériences de cette partie. Le modèle Mistral 7B, en particulier, paraît déjà suffisant pour représenter la série des modèles Mistral AI. En outre, ce modèle-ci, développé par une start-up française, pourrait effectivement être une bonne option, car son corpus d'entraînement, bien que non divulgué, comprendrait en théorie davantage de corpus français par rapport à certains modèles américains. De plus, il serait intéressant d'étendre notre étude aux LLMs purement ou largement entraînés sur des corpus français. Cela pourrait donc inclure les variantes de BERT en français, notamment CamemBERT et FlauBERT, qui sont parmi les modèles français les plus célèbres et open-source.

CamemBERT

Une équipe conjointe INRIA-Facebook a créé CamemBERT (Martin et al., 2020), un modèle BERT pré-entraîné sur 138GB de texte français. Ce modèle est basé sur

ROBERTA (Liu et al., 2019), et constitue une évolution de BERT (Devlin et al., 2019b) sur plusieurs plans, notamment par l'utilisation du masked language model comme seul objectif de pré-entrainement. Cette version amélioré nous permet de réaliser notre tâche de name-cloze. Il y a deux versions de CamemBERT : le modèle de base avec 110M de paramètres et le CamemBERT$_{\text{LARGE}}$ avec 340M de paramètres.

Les données d'entraînement du modèle sont constituées avec le sous-corpus français d'OSCAR, un corpus monolingue extrait de Common Crawl (Ortiz Suárez et al., 2020), ainsi qu'avec CCNET, un autre corpus de Common Crawl nommé CCNet (Wenzek et al., 2019), et un snapshot récent de Wikipédia française. OSCAR est composé de corpus monolingues sélectionnés par un classifieur de langue linéaire FASTTEXT (Joulin et al., 2017), tandis que CCNet est filtré à partir de Common Crawl pour être plus long et moins bruité, offrant ainsi un biais 'Wikipédia-like' en termes d'édition des textes. En outre, en utilisant le dump français officiel de Wikipédia (avril 2019), le corpus Wikipédia français prétraité avec WikiExtractor [1], fait également partie du corpus d'entraînement de ce modèle.

FlauBERT

FlauBERT est aussi un modèle de langue contextualisé (Le et al., 2020) pré-entraîné sur un corpus français hétérogène et de taille importante. FlauBERT$_{\text{BASE/LARGE}}$ présentent respectivement le nombre total de paramètres de 138 M et 373 M.

Quant aux données d'apprentissage des modèles, les chercheurs ont agrégé vingt-quatre sous-corpus de divers types, tels que Wikipedia, des livres, des données de Common Crawl, etc. Les sources principales du corpus sont les suivantes : (1) les textes monolingues issus des campagnes d'évaluation WMT19, qui se compose de quatre sous-corpus (Li et al., 2019) ; (2) les textes en français de la collection OPUS, qui comprend huit sous-corpus (Tiedemann, 2012) ; ainsi que (3) le projet Wikimedia, avec huit sous-corpus supplémentaires. La taille totale des textes intégrés, dans leur format non compressé, s'élève à 270 GB dans ce projet. Suite à un prétraitement qui inclut divers filtrages (suppression des phrases très courtes, des séquences de numéros ou d'adresses électroniques, etc.), une normalisation de l'encodage des caractères et une tokenisation utilisant l'outil Moses (Koehn et al., 2007), le corpus d'apprentissage final s'est réduit à 71 GB.

4.2.2 Items français

À partir d'environ 3 000 livres numérisés et rédigés en français, le projet French BookNLP de Lattice, auquel j'ai participé lors de mon stage, nous a mis à disposition un vaste corpus littéraire français automatiquement annoté. Grâce à une série d'outils issus de l'adaptation du projet BookNLP au français (Mélanie-Becquet et al., 2024), surtout le module de la reconnaissance des entités, nous avons pu facilement exécuter le script [2] déposé sur GitHub par l'équipe de BookNLP pour extraire des passages de livres traités par le pipeline French BookNLP. Chacun des passages contient exactement une entité PROP (une personne, PER) comme un seul *token*. Pour mieux s'adapter au français, j'ai modifié la variable de la liste '*invalid_names*', qui sert à exclure les formes familiales les plus courantes ou les tires de civilité comme entité

1. https://github.com/attardi/wikiextractor
2. https://github.com/bamman-group/gpt4-books/blob/main/generate_name_cloze/create_name_cloze_from_booknlp.py

personnelle telles que 'madame', 'mousieur','père', 'mère', etc., qui ne sont pas pertinentes pour les items de notre tâche.

L'exécution du script a permis de produire à peu près 3 000 fichiers de complétion des noms en tant que résultats initiaux. Après avoir exclu les livres dont le nombre d'éléments générés est inférieur à 100, nous sommes arrivés à conserver 2 459 livres. Cependant, limiter le nombre de livres est encore nécessaire dans le but d'éviter une durée d'expérience excessive. Nous avons identifié d'abord le nombre total de genres disponibles. Sur cette base, nous avons calculé un objectif global de 600 livres à sélectionner. Pour assurer une répartition équitable, nous avons divisé ce nombre par le nombre total de genres, ce qui nous permet d'obtenir ainsi un nombre idéal de livres à sélectionner par genre. Ainsi, nous avons finalement réussi à sélectionner de manière équilibrée 575 livres français, voir la Table 4.1. Ceux-ci couvrent les principaux genres littéraires qui sont respectivement 'roman d'aventures', 'policier', 'roman historique', 'nouvelles', 'cycles et séries', 'littérature jeunesse'. Pour tous les clozes de noms des livres, nous avons aussi procédé à une sélection aléatoire de 100 éléments pour servir à l'inférence.

Genre	**Nombre de livres**
Policier	111
Roman d'aventures	109
Littérature jeunesse	99
Roman historique	96
Cycles et séries	79
Nouvelles	78

TABLE 4.1 – Nombre de livres sélectionnés par genre

4.3 Méthode

4.3.1 Mask language modeling

En effet, les variantes françaises de BERT que nous utilisons ne peuvent pas être comparables aux Grands modèles de langue (LLMs) actuels en termes de résolution de notre tâche spécifique. Cela est dû au fait qu'elles ne prennent pas en charge le dialogue interactif avec les modèles du même type que les LLMs d'aujourd'hui. Néanmoins, CamemBERT et FlauBERT ont tous les deux adopté l'approche du masked language modeling [3]. Cette méthode implique la substitution d'une unité dans la séquence par l'unité `[MASK]`, ayant pour objectif de maximiser la probabilité de prédiction de l'unité masquée à travers la distribution de probabilité de sortie du modèle. Le masked language modeling est particulièrement adapté pour les tâches qui nécessitent une bonne compréhension contextuelle de toute la séquence. C'est pourquoi les expériences se sont effectuées sans passer par des instructions préalables mais en traitant uniquement le passage lui-même dans ce cadre-là. Par simplicité, nous avons opté pour l'utilisation du pipeline fill-mask [4] du module transformer fourni par Hugging Face. Cependant, il convient de noter que l'élément masqué est traité comme

3. https://huggingface.co/docs/transformers/tasks/masked_language_modeling#masked-language-modeling

4. https://huggingface.co/tasks/fill-mask

<mask> pour CamemBERT et comme <special1> pour FlauBERT. C'est donc dans ce contexte que nous menons notre tâche de name-cloze.

DE plus, pour CamemBERT et FlauBERT, le pipeline *fill-mask* nous a permis d'obtenir une liste des 5 substitutions les plus probables, triées par leur score de prédiction. Les résultats fournis par le pipeline comprennent un ensemble d'arguments avec leurs scores respectifs. Nous avons développé un script pour extraire et rassembler les 5 éléments `token_str` qui présentent les chaines de caractères les plus probables à marcher avec la séquence donnée pour le modèle. Ces éléments sont sauvegardés dans une nouvelle liste pour chaque extrait de sorte à s'en servir aisément pour nos analyses. Voici un exemple de structure de données retenue pour les meilleurs candidats (`token_str`) qui maximisent la probabilité de correspondre au masque (`[MASK]`) :

```
[
  {
    "score": 0.08025486767292023,
    "token": 26,
    "token_str": "un",
    "sequence": "..."
  },
  {
    "score": 0.023296814411878586,
    "token": 8639,
    "token_str": "alice",
    "sequence": "..."
  },
  ...
]
```

4.3.2 Ingénieurie multilingue des prompts

L'ingénierie multilingue des prompts est à la fois un art et une science qui consiste à créer des instructions claires et précises pour les modèles d'IA capables de comprendre et de répondre dans plusieurs langues. En fait, les LLMs à expétimenter que nous avons sélectionnés sont également transférables à plusieurs langues, y compris le français. Cependant, la capacité de communication avec les LLMs dans différentes langues n'est pas vraiment équivalente, comme le soulignent Fu et al. (2022).

C'est pourquoi nous nous sommes interrogés sur la nécessité de reformuler ou de traduire les instructions en français, étant donné que nous allons fournir un extrait du livre français à prédire par le modèle. Pour ce faire, nous avons conçu deux versions de prompts pour un mini-test. La première est un prompt purement en français qui comprend à la fois les instructions et les exemples fournis. Par exemple, nous avons choisi une phrase courte contenant un nom de personnage de l'œuvre *Madame Bovary* et une autre issue de *Notre-Dame de Paris*. La seconde version utilise le modèle de prompt de Chang et al. (2023) que l'on a déjà montré dans le chapitre 2, voir l'image 2.4, tout en anglais. Après avoir terminé le mini-test pour évaluer l'efficacité des deux versions de prompts, nous avons observé que, pour plusieurs livres populaires testés, le prompt original de Chang et al. (2023) a conduit à des performances plus élevées des LLM par rapport à la version en français. Il est donc préférable d'utiliser le prompt en anglais. Il semble donc raisonnable de déduire que les LLM

choisis ne sont pas encore en mesure de se généraliser pleinement à d'autres langues pour les instructions. Toutefois, compte tenu de l'apprentissage à faible échantillonnage (few-shot learning), les exemples fournis jouent également un rôle clé dans la performance du modèle pour notre expérience. Bien que cette déduction ne soit peut-être pas pour autant assez convaincante, nous avons finalement décidé d'utiliser le prompt en anglais.

4.4 Résultats

Les expériences présentées dans cette section ont été menées sur quatre modèles de langue : FlauBERT, CamemBERT, OLMo 7B et Mistral 7B. Pour FlauBERT, nous avons évalué les impacts de la taille des paramètres en testant simultanément $\text{FlauBERT}_{\text{BASE}}$ et $\text{FlauBERT}_{\text{LARGE}}$, pour comparer les performances entre les modèles BERT de différentes tailles.

En tout cas, les résultats obtenus sont toujours beaucoup inférieurs à ceux obtenus en anglais, avec des scores relativement faibles. Nous allons ensuite, dans cette partie, analyser les résultats de l'adaptation en français à travers trois axes principaux : la comparaison de la distribution des scores moyens, l'analyse des scores moyens par genre littéraire et l'analyse des scores moyens en fonction du droit d'auteur.

4.4.1 Analyse comparative des scores en moyenne

Quant à l'évaluation de la performance des modèles, nous avons adopté deux approches distinctes pour interpréter les résultats obtenus avec $\text{FlauBERT}_{\text{LARGE}}$ et $\text{CamemBERT}_{\text{LARGE}}$. Lorsque nous notons $\text{FlauBERT}_{\text{LARGE}}[0]$ ou $\text{CamemBERT}_{\text{LARGE}}[0]$, cela signifie que seul le meilleur candidat parmi les token_strings est retenu comme prédiction finale du modèle. En d'autres termes, nous considérons que le modèle a généré la prédiction la plus pertinente lorsqu'il s'agit du premier élément de la liste des candidats. Par contre, si nous évaluons directement les performances de $\text{CamemBERT}_{\text{LARGE}}$ ou $\text{FlauBERT}_{\text{LARGE}}$, sans la notation [0], nous obtenons une liste de 5 éléments considérés comme des prédictions potentielles du modèle. Dans cette approche, le résultat 'gold' (ou la réponse correcte) est inclus dans cette liste. Si le résultat gold se trouve bien dans cette liste, cela indique que les modèles ont réussi à le prédire avec succès.

Afin d'analyser et de visualiser les performances des modèles sur les items en français, nous avons d'abord calculé les métriques d'évaluation standards à l'aide de la fonction `describe()` du package Pandas. A l'aide de ces métriques, un boxplot associé a été généré pour illustrer la distribution des scores obtenus par différents modèles, voir la Figure 4.1.

L'analyse des résultats des métriques met en évidence que la précision de Camembert Large se démarque avec une moyenne la plus élevée (0,0290). Cependant, une variation significative est observée, comme en témoigne un écart type de 0,0477, ce qui indique des performances inégales, aussi visibles à travers une distribution plus large et une présence d'écarts plus importants dans la figure concernée.

Par contre, Olmo 7B *accuracy* et Flaubert Base *accuracy*, ainsi que Flaubert Large[0] *accuracy*, ont des moyennes généralement plus faibles, avec des médianes proches de zéro, ce qui suggère que ces modèles n'ont pas bien performé dans la plu-

part des cas. On a donc pu visualiser les distributions de ces trois modèles très serrées autour de zéro.

CamemBERT se montre beaucoup plus performant dans notre tâche par rapport à FlauBERT. Cependant, Olmo 7B semble éprouver des difficultés pour prédire les éléments français. Pour Mistral 7B, les résultats sont quelque peu similaires à ceux de CamemBERT Large[0] et ils sont encore supérieurs à ceux du modèle OLMo 7B. Enfin, il est également vérifié que les modèles de plus grande taille ont des performances globalement meilleures, car les BERTS Large présente toujours le meilleur score par rapport à celui de base.

Ensuite , nous nous intéressons à analyser les valeurs aberrants pour le résultat de Cammenbert Large Accuracy. Les top 20 valeurs extrêmes des prédictions du modèle, sont illustrés dans le table 4.2.

doc_name	genre	Camembert Large Accuracy
1891_Witt-Henriette-de_La_Petite-fille-aux-grains-d'or	littérature jeunesse	0.49
1843_Woillez-Catherine_Leontine-et-Marie	littérature jeunesse	0.46
1861_Villars-Fanny_Marie-et-Marguerite	roman historique	0.31
1853_Sue-Eugene_Les-Mysteres-du-peuple_Tome-IX	roman d'aventures	0.27
1888_Beaugrand-Honore_Jeanne-la-fileuse	roman historique	0.27
1878_Gouraud-Julie_Cousine-Marie	littérature jeunesse	0.26
2008_Lemaitre-Pierre_Robe-de-Marie	policier	0.23
1885_Amero-Constant_Le-tour-de-France-d'un-petit-garçon	littérature jeunesse	0.22
1890_Hameau-Louise_Mademoiselle-Pourquoi	littérature jeunesse	0.21
1867_Gouraud-Julie_Le-petit-colporteur	littérature jeunesse	0.20
1852_Carraud-Zulma-Tourangin-Mme_La-petite-Jeanne	littérature jeunesse	0.19
1854_Barbey-d-Aurevilly-Jules_L-ensorcelee	roman historique	0.18
1942_Camus-Albert_L-etranger	-	0.17
1908_Dhanys-Marcel_Le-Roman-du-genie	littérature jeunesse	0.15
1873_Viollet-le-Duc-Eugene_Histoire-d'un-chateau	littérature jeunesse	0.14
1847_Collin-de-Plancy-Jacques_Legende-du-Juif-Errant	roman historique	0.14
1845_Dumas-Alexandre_Histoire-d'un-Casse-noisette	littérature jeunesse	0.14
2003_Grange-Jean-Christophe_L-empire-des-loups	policier	0.14
1846_Balzac-Honore-de_Les-Chouans	roman historique	0.13
1838_Dupin-Antoinette_Comment-tout-finit_Tome-2	nouvelles	0.13

TABLE 4.2 – Valeurs aberrantes pour le modèle Camembert Large Accuracy

On remarque que la majorité des valeurs aberrantes appartiennent à ce genre, ce qui pourrait indiquer que les caractéristiques de ce type de littérature entraînent une variabilité plus élevée dans les évaluations de précision par le modèle Camembert Large. 1843_Woillez-Catherine_Leontine-et-Marie (0.46) et 1891_Witt-Henriette-de_La-Petite-fille-aux-grains-d'or (0.49) ont des scores de précision nettement plus élevés que les autres.

4.4.2 Analyse des moyennes par genre littéraire

Cette sous-section est dédiée à l'exploration des performances de différents modèles de langue sur les six principaux genres littéraires que nous avons fixés précédemment : cycles et séries, littérature jeunesse, nouvelles, romans policiers, romans d'aventures et romans historiques. Les résultats sont présentés sous la forme d'un diagramme à barres montrant les précisions moyennes par genre littéraire pour chaque modèle, comme illustré à la Figure 4.2.

La littérature jeunesse et les romans historiques tendent à avoir les plus hautes précisions moyennes pour la plupart des modèles, notamment Camembert Large et Camembert Large[0]. À l'inverse, le genre policier se distingue par des précisions

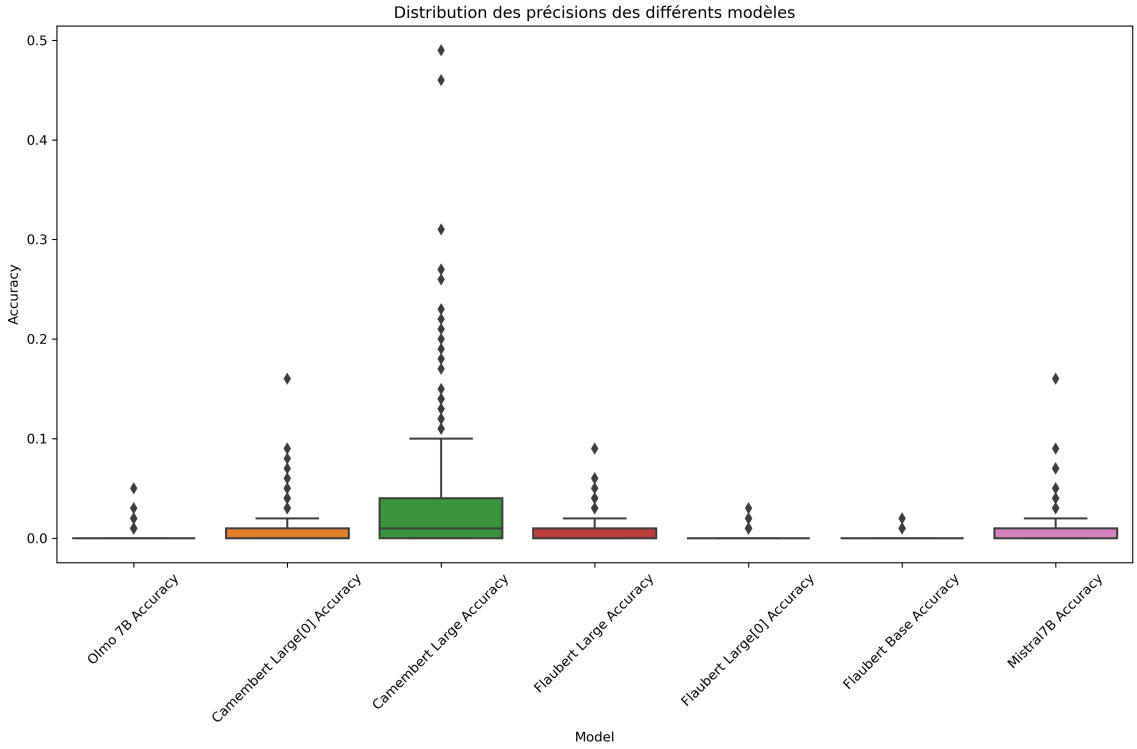


FIGURE 4.1 – Distribution des scores obtenus dans les expériences françaises

moyennes relativement basses sur la majorité des modèles, ce qui pourrait suggérer que les textes de ce type posent plus de défis aux modèles.

En général, Camembert Large manifeste une variation notable dans les précisions moyennes entre les genres, tandis que d'autres modèles comme Olmo 7B et Flaubert Base affichent des précisions moyennes plus faibles et plus uniformes. Par ailleurs, Mistral 7B ne montre pas une grande variabilité dans les scores moyens en fonction du genre littéraire, bien que ceux-ci restent toujours en deçà de 0,01.

4.4.3 Analyse des moyennes par droit d'auteur

Parallèlement, conformément à la législation sur la propriété intellectuelle en France, abordée en premier chapitre, nous avons déterminé le statut du droit d'auteur pour chaque livre en se basant sur la date de décès de l'auteur. Les résultats sont présentés sous forme de graphiques à barres à la Figure 4.3, en mettant en évidence les précisions moyennes des différents statuts de droit d'auteur pour chaque modèle.

Globalement, les ouvrages dans le domaine public se révèlent avoir des précisions moyennes légèrement plus élevées pour la plupart des modèles, notamment Camembert Large et Mistral 7B. En revanche, les livres soumis au droit d'auteur affichent des précisions moyennes inférieures par rapport aux œuvres du domaine public pour la quasi-totalité des modèles. Cependant, seul le modèle Camembert Large démontre une variation des précisions entre les deux statuts juridiques jugée suffisamment significative. En résumé, la différence entre les ouvrages soumis au droit d'auteur et ceux qui ne le sont pas est moins marquée en français qu'en anglais. Cette observation pourrait s'expliquer par la non-supériorité de tous les scores dans les résultats français, rendant ainsi leur comparaison plus difficile.

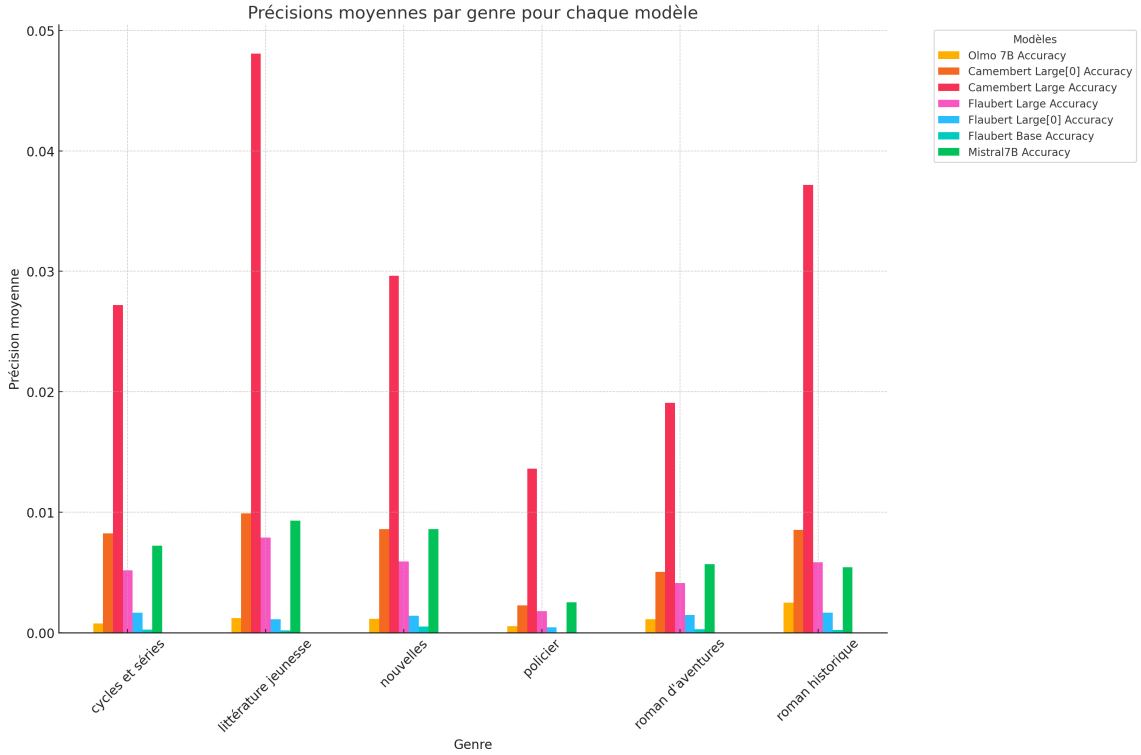


FIGURE 4.2 – Moyenne des exactitudes pour les livres français de divers genres

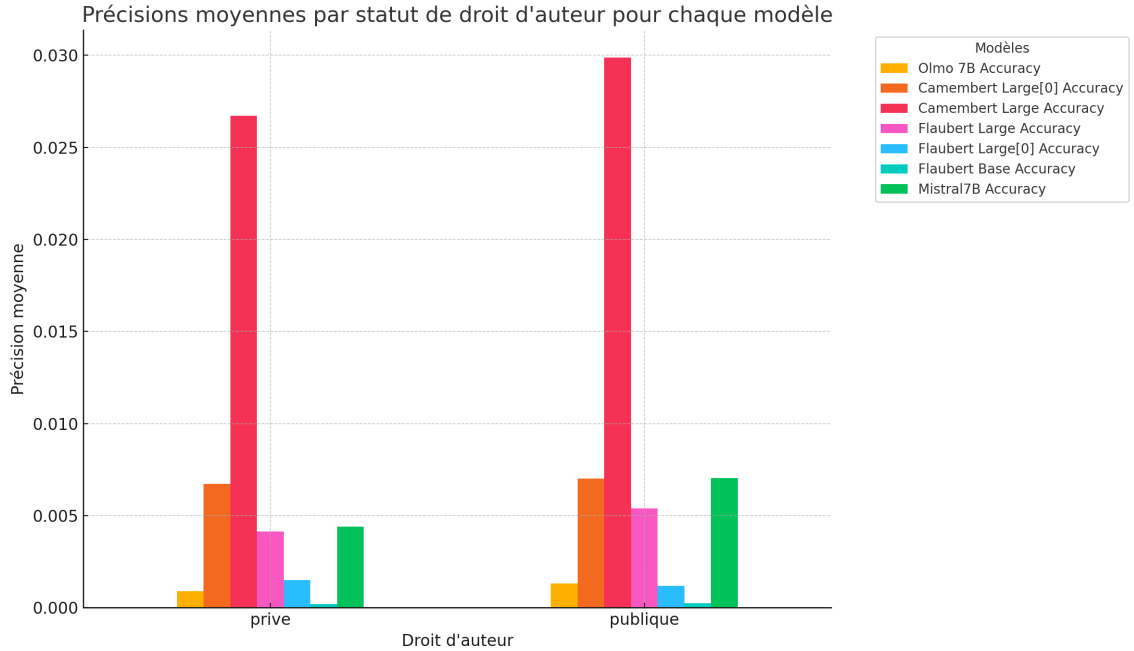


FIGURE 4.3 – Moyenne des exactitudes pour les livres français publiques et privés

4.5 Discussion

Nous avons reproduit les expériences de Chang et al. (2023), en nous généralisant à la langue française. Nos résultats révèlent que les modèles français basés sur des variantes de BERT sont particulièrement performants. Concernant les Grands modèles de langue (LLM) classiques tels qu'Olmo 7B et Mistral 7B, ils ont présenté des scores moyens toujours faibles. Il se pose la question de dire que les modèles ont mémorisé les livres français, indépendamment de leur statut d'auteur, car les moyennes du modèle Mistral 7B sont similaires en anglais et en français. Ainsi, lorsqu'on parle de la capacité de généralisation dans nos expériences en français, plusieurs problèmes demeurent à appréhender.

4.5.1 Limites du Booknlp français

L'objectif principal du French BookNLP est de réutiliser au maximum les outils existants, initialement conçus pour le corpus littéraire anglais, et de les adapter au français. Cependant, en procédant à cette extension, nous avons identifié plusieurs points d'amélioration, notamment en ce qui concerne les systèmes d'annotation, particulièrement en raison des spécificités de la langue française. Ce constat n'est normalement pas surprenant puisque le projet n'a été lancé que récemment au laboratoire Lattice et qu'il est en cours d'optimisation concernant les annotations spécifiques.

Parmi ces points d'amélioration, notons l'annotation incorrecte ou non sophistiquée qui pourrait affecter les résultats de nos expériences. Cela se pose d'emblée lors de la constitution des items français. En effet, nous avons utilisé le script de Chang et al. (2023) pour masquer un nom de personnage dans un extrait français par le biais du pipeline des entités et des tokens reconnus, ce qui a entraîné des problèmes. Par exemple, dans l'annotation des drames français, où le format de dialogue est courant, les noms propres suivis d'un tiret (-) ont été toujours mal reconnus dans le fichier .entities. Un nom propre combiné avec .— est traité comme un seul token, mais n'est pas identifié comme un nom de personnage nommé, ce qui le laisse inchangé dans le passage. Cela conduit à des cas où l'élément masqué, comme le nom ALCIDE dans l'exemple fourni [5], est directement exposé, ce qui va certainement fausser ou biaiser la nature prédictive de l'expérience. Hors tout cela, nous avons remarqué que le problème d'encodage des données en français persiste dans les chaînes de traitement actuelles.

4.5.2 Nécessité de Multiples Chances de Prédiction

Dans le cadre de notre étude en français, nous avons envisagé deux approches d'interprétation pour le comptage des scores d'*accuracy*, notamment pour les modèles BERT français. Il convient de prêter davantage d'attention au fait que nous avons examiné les résultats du Top 5, qui ont significativement amélioré les performances. Cependant, il est important de souligner que, dans la plupart des cas, le score d'accuracy est calculé en ne considérant que le premier candidat qui présente la plus grande probabilité de prédiction du modèle.

Cela nous amène à nous interroger sur la pertinence d'offrir plusieurs chances de prédiction au modèle. Il est possible que la mémorisation par le modèle pour cer-

5. cf. on lui en montra plusieurs en argent. «Des montres d'or, dit [MASK] en repoussant avec mépris celles d'argent. —Tu es donc devenu bien riche ? répondit le cousin. ALCIDE.—Oui ; on nous a donné de quoi acheter des montres en or.

taines informations très précises soit floue, ce qui est bien justifié. En effet, pour un humain, après avoir lu un livre, on est normalement capable de mémoriser les plots essentiels de l'œuvre, même si cela implique de nombreux personnages. Cependant, pour un seul extrait à prédire, il peut être très difficile de donner immédiatement et avec précision qui est le bon personnage ici. Cette capacité à mémoriser et à prédire soulève des questions sur la nature même de la mémorisation. D'un côté, on pourrait envisager que la mémorisation est une sorte d'atmosphère tracée dans la mémoire, qui pourrait contribuer à deviner le bon résultat. Ou bien, cela pourrait être interprété comme la capacité de mémoriser de manière très concrète, permettant ainsi de répondre rapidement aux questions connexes. Donc, nous avons la peine de différencier les deux notions connexes dans ce cas-là : Assimilation et Mémorisation. On ne va pas entrer en très détail ici.

En outre, pour les Grands modèles de langue (LLM) d'aujourd'hui, nous paramétrons toujours [0] pour fixer la meuilleure réponse comme la suite des dialogues lors du prompting. Néanmoins, il est tout à fait possible de générer plusieurs réponses. Ainsi, ces différentes approches nous incitent à reconsidérer la méthodologie originale et à explorer d'autres méthodes pour améliorer la performance et la pertinence des prédictions dans le futur.

4.6 Conclusion

Ce chapitre est consacré à l'application de la méthodologie des expériences de Chang et al. (2023) dans le contexte de la langue française. Après avoir modifié les listes d'invalidités, qui déterminent les noms de personnages à masquer, et une sélection équilibrée, nous avons généré aléatoirement 57 500 items pour 575 ouvrages pour notre expérience en français.

Les résultats, évalués également à l'aide de la métrique d'accuracy des clozes de nom, sont généralement bas. Cependant, les variantes de BERT principalement entraînées sur le français, à savoir CamemBERT et FlauBERT, ont présenté des scores relativement élevés en moyenne. Les grands modèles de langue (LLM) sélectionnés, OLMo 7B et Mistral 7B, n'ont pas apporté de grande variation, en termes de moyenne des scores, lorsqu'ils sont passés de l'anglais au français.

L'analyse a principalement porté sur deux facteurs clés : la catégorie littéraire et le statut du droit d'auteur. Concernant le genre littéraire, la littérature jeunesse domine pour tous les modèles en termes de score d'accuracy moyen. Cependant, la différence entre les livres du domaine public et ceux du domaine privé n'est pas nette pour tous les modèles, bien que légèrement inférieure. Cette différence n'est pas très évidente en français.

Nous avons également examiné les limites et procédé aux considérations suivantes les expériences en français. D'une part, en réutilisant l'outil de génération de name-cloze spécifiquement destiné à l'anglais sur un corpus annoté en français moins parfait, il est nécessaire d'être prudent dans l'interprétation des résultats. Certaines annotations incorrectes peuvent être considérées comme des fraudes et peuvent gravement altérer le résultat final. D'autre part, la méthodologie pour obtenir des prédictions pourrait impliquer plusieurs tentatives, ce qui soulève une question légitime, en particulier compte tenu de l'amélioration significative des scores de CamemBERT Large [0] et CamemBERT Large (Top 5).

En résumé, les expériences en français ont produit des résultats plutôt bas. Néanmoins, les LLM ont plus ou moins démontré la capacité de se généraliser au français,

étant donné que les moyennes de scores sont similaires à celles des expériences en anglais. Ce chapitre a également permis de reconsidérer la méthodologie initiale des expériences de Chang et al. (2023), en se questionnant sur la nécessité d'accepter un certain nombre de résultats comme des prédictions valides.

CHAPITRE 5

CONCLUSION

5.1 Synthèse des travaux

Nous avons étudié dans ce mémoire la mémorisation des textes littéraires par des grands modèles de langue en répliquant et en étendant les expériences de complétion de nom (*name cloze*) de Chang et al. (2023). Nous avons traité cette problématique à travers trois axes principaux : la réplication des expériences initiales avec des modèles libres d'accès ; l'implémentation d'expériences supplémentaires, nous concentrant notamment sur les corpus de pré-entraînement, les points d'arrêt d'apprentissage et des variantes de la complétion de nom ; et finalement l'application de cette méthodologie à la littérature française.

Dans le chapitre 2, nous avons répliqué les résultats de Chang et al. (2023) en utilisant des modèles ouverts pour prédire des noms propres dans des extraits littéraires. Nos résultats montrent que les modèles ouverts sont moins performants sur la tâche que ceux d'OpenAI : GPT-4 en particulier demeure le plus performant. L'influence du statut des droits d'auteur, du genre littéraire et de la popularité des livre sur l'Internet s'est montrée — comme dans l'étude de Chang et al. (2023) — déterminante pour la performance. Appartenir au domaine public, être un livre de science-fiction et avoir beaucoup de *hits* en ligne augmentent la probabilité d'un score élevé.

Dans le chapitre 3, nous avons évalué de manière empirique la capacité de mémorisation de grands modèles de langue. En reconnaissant les limites d'une seule métrique d'évaluation, nous avons mené d'autres expériences pour comprendre le processus de mémorisation et son impact sur les performances des modèles, telles que le trajectoire de mémorisation au fil de l'évolution des points d'arrêt d'apprentissage, la statistique des personnages principaux et secondaires, etc. Dans les expériences de trajectoire, nous avons constaté que, par rapport aux livres publics, les livres protégés par le droit d'auteur présentent généralement des fluctuations à différents points d'arrêt et un taux d'oubli relativement élevé. En outre, OLMo7B est le plus performant pour prédire les protagonistes de romans. Nous avons également vérifié la faisabilité de l'évaluation des textes littéralement mémorisés en tant que tâche en aval. Après toutes les expériences des vérifications empiriques, nous avons appris que les modèles ont effectivement tendance à mémoriser les livres populaires, ce qui se manifeste aussi dans la tâche en aval de la prédiction des séquences continues.

Dans le chapitre 4, nous avons adapté la méthodologie de Chang et al. (2023) à la langue française. Les résultats obtenus avec des modèles tels que CamemBERT et FlauBERT, qui sont principalement entraînés sur des données françaises étaient plus

élevés que ceux des autres modèles comptant pourtant beaucoup plus de paramètres. Les modèles OLMo7B et Mistral7B n'ont pas montré de grande variation dans leurs performances entre l'anglais et le français. L'analyse des facteurs clés tels que le genre littéraire et le statut du droit d'auteur a révélé des tendances similaires à celles observées dans les expériences en anglais, bien que moins marquées.

5.2 Limitations

Dans cette section, nous évoquerons plusieurs limitations à adresser pour notre travail de recherche.

5.2.1 Problèmes des données (p.e. opacité et contamination)

L'opacité des données utilisées pour les modèles testés rend la validation des résultats difficile. À l'exception du modèle OLMo, les données derrière les modèles de notre expérience sont fondamentalement inconnues. Même avec de brèves documentations, nous avons des difficultés à déterminer si un tel livre existe réellement dans les données d'entraînement. Les résultats montrent que certains livres soumis au droit d'auteur, qui ne devraient normalement pas être utilisés pour l'entraînement des modèles, ont obtenu des scores relativement élevés. Cependant, nous ne pouvons pas, à partir de ces scores, conclure définitivement qu'un modèle s'est entraîné avec des livres du domaine privé. D'autre part, la présence des données d'Internet risque de fausser l'interprétation des résultats. À partir de l'expérience décrite dans la section 3.1.3, nous pouvons confirmer que les données d'entraînement sont contaminées, puisque certains extraits de notre jeu d'évaluation ont été implicitement incorporé dans le corpus d'entraînement, principalement par les données Internet. Cela rend les résultats finalement moins convaincants.

5.2.2 Problèmes des méthodes

Dans notre travail, l'application de la méthode de fouille des données par l'intermédiaire de l'inférence de noms (name cloze inference) a présenté certaines limitations. Les résultats obtenus avec cette méthode ont été globalement peu convaincants, qui ne permettent pas une analyse approfondie par exemple dans l'expérience 3.1.3. En particulier, les modèles Llama, Pythia7B et Pythia13B ont démontré des performances décevantes avec des scores constamment très bas. De plus, la connaissance inégale des modèles aux livres spécifiques pourrait aussi fausser l'interprétation des résultats. En se référant aux scores obtenus, pour quelques livres spécifiques, tels que *Alice in Wonderland*, *Harry Potter*, ou *Pride and Prejudice*, etc, les modèles ont eu une mémorisation beaucoup plus forte par rapport à un grand nombre d'autres. Enfin, la réputation des auteurs pourrait également impacter la performance. Lorsque des auteurs bien connus, comme Zola, ont tendance à créer un univers littéraire cohérent et à utiliser le même nom de personnage à travers leurs différentes œuvres, la prédiction des noms de personnages par les modèles n'est pas tout à fait aléatoire. Les modèles pourraient en fait se référer à l'univers spécifique de ces auteurs pour faire des prédictions, ce qui est moins juste.

5.3 Perspectives

Les travaux menés dans ce mémoire peuvent être poursuivis dans différentes directions.

5.3.1 Évaluation des méthodes

D'un côté, nous pourrions peaufiner les méthodes d'évaluation. Nous nous intéresserions à quantifier l'impact de la contamination des données d'entraînement sur les performances des modèles, surtout dans le cas des modèles opaques. Au lieu de se contenter de la moyenne des indexes pour l'expérience de trajectoire de mémorisation, nous pourrions développer un nouvel algorithme de score de stabilité de mémorisation. Ce score évaluerait la variation des indexes à travers les points d'arrêt successifs, permettant de tracer une courbe de stabilité. En superposant cette courbe à celle de performance d'*accuracy*, nous pourrions tenter d'identifier les phases d'entraînement correspondant à la mémorisation d'un livre entier par le modèle. Qualitativement, nous voulons également étudier en plus de profondeur l'impact que les différentes natures de texte ont sur le comportement d'apprentissage : si un roman intégral donne un meilleur résultat que des textes segmentés, s'il existe une différence quant à la mémorisation des œuvres littéraires et les textes non littéraires, ou encore, entre différentes genres de romans. Nous pourrions concevoir de nombreux expériences de *probing* sur ces sujets à l'intersection de TAL et de l'analyse culturelle.

5.3.2 Atténuation des biais

D'un autre côté, les biais langagiers et littéraires observés dans nos expériences reflète l'image de la prédominance culturelle. Ce phénomène pose des défis importants que nous devons surmonter pour assurer une représentation équitable. Étant donné que notre application à la langue française n'a pas été aussi réussie que prévu, car les textes en anglais prédominent normalement dans le corpus. Pour équilibrer les corpus d'entraînement multilingues tout en maintenant la robustesse du modèle, il semble intéressant de considérer des stratégies telles que l'ajustement de la pondération des couches du modèle. Nous voulons aussi examiner davantage la bipolarisation observée entre certains livres très populaires, qui ont connu un taux de succès surprenant, et la majorité des autres. En effet, nous constatons une fracture de la mémorisation drastique, qui peut être attribué à des biais littéraires et un seuil potentiel de popularité qui revalorise la pondération. Ces facteurs ont un impact significatif sur les sorties des modèles, qui ont tendance à réutiliser la narration de ces livres populaires peu nombreux. Le travail futur serait d'évaluer en détail l'apparition de ce phénomène et de proposer des solutions pour améliorer la diversité langagière et l'équité culturelle pour les LLMs.

BIBLIOGRAPHIE

AI@Meta (2024). Llama 3 model card. – Cité page 8.

Anthropic (2024). Introducing Claude. – Cité page 8.

Bamman, D. (2021). BookNLP. – Cité page 16.

Bamman, D., Lewke, O., and Mansoor, A. (2020). An annotated dataset of coreference in English literature. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association. – Cité page 11.

Biderman, S., PRASHANTH, U., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. (2024). Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36. – Cité pages 8 et 27.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR. – Cité pages 14 et 27.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. (2023). Extracting training data from diffusion models. – Cité page 34.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., and Zhang, C. (2022). Quantifying memorization across neural language models. *ArXiv*, abs/2202.07646. – Cité page 8.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models. – Cité page 39.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D. X., Erlingsson, Ú., Oprea, A., and Raffel, C. (2020). Extracting training data from large language models. In *USENIX Security Symposium*. – Cité page 8.

Chang, K., Cramer, M., Soni, S., and Bamman, D. (2023). Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics. – Cité pages 4, 5, 9, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 32, 34, 36, 37, 38, 40, 43, 46, 51, 52, 53 et 55.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357. – Cité page 36.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. – Cité page 8.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. – Cité page 44.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. – Cité page 22.

Elazar, Y., Bhagia, A., Magnusson, I., Ravichander, A., Schwenk, D., Suhr, A., Walsh, P., Groeneveld, D., Soldaini, L., Singh, S., et al. (2023). What's in my big data? *arXiv preprint arXiv:2310.20707*. – Cité page 14.

Fu, Z., Zhou, W., Xu, J., Zhou, H., and Li, L. (2022). Contextual representation learning beyond masked language modeling. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2701–2714, Dublin, Ireland. Association for Computational Linguistics. – Cité page 46.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*. – Cité page 14.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., et al. (2024). Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*. – Cité pages 8, 13 et 28.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. (2023). Foundation models and fair use. *ArXiv*, abs/2303.15715. – Cité page 9.

Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695. – Cité page 15.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. – Cité page 37.

Hui, B., Yang, Y., Yuan, H., Burlina, P., Gong, N. Z., and Cao, Y. (2021). Practical blind membership inference attack via differential comparisons. *ArXiv*, abs/2101.01341. – Cité page 8.

IBM (2024). Qu'est-ce qu'un grand modèle de langage ? | IBM. – Cité page 8.

Jia, J., Salem, A., Backes, M., Zhang, Y., and Gong, N. Z. (2019). Memguard: Defending against black-box membership inference attacks via adversarial examples. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. – Cité page 9.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*. – Cité page 12.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics. – Cité page 44.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Ananiadou, S., editor, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics. – Cité page 44.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. arXiv:1912.05372 [cs]. – Cité page 44.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. (2022). Deduplicating training data makes language models better. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics. – Cité page 8.

Li, B., Li, Y., Xu, C., Lin, Y., Liu, J., Liu, H., Wang, Z., Zhang, Y., Xu, N., Wang, Z., Feng, K., Chen, H., Liu, T., Li, Y., Wang, Q., Xiao, T., and Zhu, J. (2019). The NiuTrans machine translation systems for WMT19. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K., editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266, Florence, Italy. Association for Computational Linguistics. – Cité page 44.

Liu, J., Min, S., Zettlemoyer, L., Choi, Y., and Hajishirzi, H. (2024). Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*. – Cité page 34.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. – Cité page 44.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, V., Seddah, D., and Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219. arXiv:1911.03894 [cs]. – Cité page 43.

Mélanie-Becquet, F., Barré, J., Seminck, O., Plancq, C., Naguib, M., Pastor, M., and Poibeau, T. (2024). Booknlp-fr, the french versant of booknlp. a tailored pipeline for 19th and 20th century french literature. – Cité page 44.

Onishi, T., Wang, H., Bansal, M., Gimpel, K., and McAllester, D. (2016). Who did what: A large-scale person-centered cloze dataset. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Austin, Texas. Association for Computational Linguistics. – Cité page 15.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., and al. (2024). Gpt-4 technical report. – Cité page 8.

Ortiz Suárez, P. J., Romary, L., and Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics. – Cité page 44.

Pang, J., Ye, F., Wang, L., Yu, D., Wong, D. F., Shi, S., and Tu, Z. (2024). Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models. arXiv:2401.08350 [cs]. – Cité page 22.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. – Cité page 13.

Rahman, M., Rahman, T., Laganière, R., and Mohammed, N. (2018). Membership inference attack against differentially private deep learning model. *Trans. Data Priv.*, 11:61–79. – Cité page 9.

Salem, A., Zhang, Y., Humbert, M., Fritz, M., and Backes, M. (2018). Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *ArXiv*, abs/1806.01246. – Cité page 8.

Sanyal, A., Hu, Y., and Yang, F. (2022). How unfair is private learning? In *Uncertainty in Artificial Intelligence*, pages 1738–1748. PMLR. – Cité page 8.

Shazeer, N. (2020). Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*. – Cité page 15.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2016). Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. – Cité page 9.

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., et al. (2024). Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*. – Cité pages 13, 14 et 28.

Staab, R., Vero, M., Balunović, M., and Vechev, M. (2024). Beyond memorization: Violating privacy via inference with large language models. – Cité page 8.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. (2024). Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063. – Cité page 15.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA). – Cité page 44.

Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. (2022a). Memorization without overfitting: Analyzing the training dynamics of large language models. *ArXiv*, abs/2205.10770. – Cité page 9.

Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. (2022b). Memorization without overfitting: Analyzing the training dynamics of large language models. – Cité page 31.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). Llama: Open and efficient foundation language models. – Cité page 12.

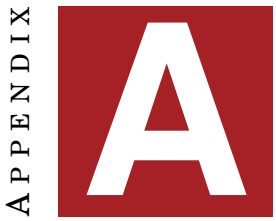
Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. – Cité pages 12 et 15.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need. – Cité pages 8, 12 et 13.

Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K. R., Wadden, D., MacMillan, K., Smith, N. A., Beltagy, I., and Hajishirzi, H. (2023). How far can camels go? exploring the state of instruction tuning on open resources. – Cité page 13.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. (2019). Ccnet: Extracting high quality monolingual datasets from web crawl data. – Cité page 44.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115. – Cité page 8.



ANNEXE

A.1 Prédiction d'OLMo à différents checkpoints

A.1.1 The Silmarillion

Nom	OriginalIndex	Gold	Item
Sauron	28	Sauron	But for a long time he did not dare to challenge the Lords of the Sea, and he withdrew from the c...
Checkpoint: step50000			
Sauron	39	Sauron	And one by one, sooner or later, according to their native strength and to the good or evil of th...
Checkpoint: step100000			
Sauron	28	Sauron	But for a long time he did not dare to challenge the Lords of the Sea, and he withdrew from the c...
Checkpoint: step150000			
Sauron	28	Sauron	But for a long time he did not dare to challenge the Lords of the Sea, and he withdrew from the c...
Checkpoint: step250000			
Sauron	56	Sauron	For now, having the ears of men, [MASK] with many arguments gainsaid all that the Valar had taugh...
Checkpoint: step300000			
Sauron	28	Sauron	But for a long time he did not dare to challenge the Lords of the Sea, and he withdrew from the c...
Sauron	39	Sauron	And one by one, sooner or later, according to their native strength and to the good or evil of th...
Checkpoint: step350000			
Sauron	39	Sauron	And one by one, sooner or later, according to their native strength and to the good or evil of th...
Checkpoint: step450000			

Sauron	28	Sauron	But for a long time he did not dare to challenge the Lords of the Sea, and he withdrew from the c...
Sauron	39	Sauron	And one by one, sooner or later, according to their native strength and to the good or evil of th...
Morgoth	95	Morgoth	Certain it is whence they came, and the evil truth was enhanced and poisoned by lies; but the Sin...
Checkpoint: step500000			
Sauron	28	Sauron	But for a long time he did not dare to challenge the Lords of the Sea, and he withdrew from the c...
Sauron	38	Sauron	But the years passed, and the King felt the shadow of death approach, as his days lengthened; and...
Sauron	39	Sauron	And one by one, sooner or later, according to their native strength and to the good or evil of th...
Checkpoint: step557000			

A.1.2 The Chosen

Nom	**OriginalIndex**	**Gold**	**Item**
Danny	12	Danny	I'd much rather be pressured. He's a nice kid, though. Your sister's pretty nice, too, I said....
Checkpoint: step400000			
Danny	54	Danny	I knew he would find out about it sooner or later, he said softly, looking very sad. I hope yo...
Danny	90	Danny	To hell with you and your damn silence. By the time the fall semester officially began two days l...
Checkpoint: step450000			
Danny	56	Danny	But he was very carefully avoiding me, and I knew enough to stay away from him. I didn't want wor...
Danny	92	Danny	I looked at him. I thought he might be sick. I thought your sister said He's not sick, [MASK...
Danny	99	Danny	And he went from the room, leaving me as bewildered as I had been before. I had planned to talk t...
Checkpoint: step557000			

A.1.3 The Mystery of Udolpho

Nom	OriginalIndex	Gold	Item
Emily	93	Emily	Would you not otherwise be willing to hope for my reformation–and could you bear, by estranging ...
Checkpoint: step250000			
Emily	64	Emily	The holy father began the service, and [MASK] again commanded her feelings, till the coffin was l...
Emily	71	Emily	It was open upon the table, before her, among some loose drawings, having, with them, been taken ...
Emily	93	Emily	Would you not otherwise be willing to hope for my reformation–and could you bear, by estranging ...
Checkpoint: step350000			
Emily	93	Emily	Would you not otherwise be willing to hope for my reformation–and could you bear, by estranging ...
Checkpoint: step400000			
Emily	93	Emily	Would you not otherwise be willing to hope for my reformation–and could you bear, by estranging ...
Checkpoint: step450000			
Emily	6	Emily	[MASK] sat by her father, holding his hand, and, while she listened to the old man, her heart swe...
Emily	17	Emily	[MASK] smiled through her tears upon her father: 'Dear sir,' said she, and her voice trembled; sh...
Emily	62	Emily	At the sound of his voice, [MASK] turned her eyes, and a gleam of recollection seemed to shoot at...
Emily	71	Emily	It was open upon the table, before her, among some loose drawings, having, with them, been taken ...
Emily	93	Emily	Would you not otherwise be willing to hope for my reformation–and could you bear, by estranging ...
Checkpoint: step500000			
Emily	6	Emily	[MASK] sat by her father, holding his hand, and, while she listened to the old man, her heart swe...
Emily	17	Emily	[MASK] smiled through her tears upon her father: 'Dear sir,' said she, and her voice trembled; sh...

Emily	62	Emily	At the sound of his voice, [MASK] turned her eyes, and a gleam of recollection seemed to shoot at...
Emily	64	Emily	The holy father began the service, and [MASK] again commanded her feelings, till the coffin was l...
Emily	71	Emily	It was open upon the table, before her, among some loose drawings, having, with them, been taken ...
Emily	93	Emily	Would you not otherwise be willing to hope for my reformation–and could you bear, by estranging ...
Checkpoint: step557000			

A.1.4 Pride and Prejudice

Nom	OriginalIndex	Gold	Item
Elizabeth	3	Elizabeth	I would go and see her if I could have the carriage. [MASK], feeling really anxious, was determi...
Jane	62	Jane	On this point she was soon satisfied; and two or three little circumstances occurred ere they par...
Checkpoint: step400000			
Jane	79	Jane	She anticipated what would be felt in the family when her situation became known; she was aware t...
Checkpoint: step450000			
Elizabeth	3	Elizabeth	I would go and see her if I could have the carriage. [MASK], feeling really anxious, was determi...
Elizabeth	10	Elizabeth	At any rate, she cannot grow many degrees worse, without authorising us to lock her up for the re...
Elizabeth	12	Elizabeth	[MASK] could safely say that it was a great happiness where that was the case, and with equal sin...
Elizabeth	19	Elizabeth	[MASK], who had a letter to write, went into the breakfast room for that purpose soon after tea; ...
Elizabeth	20	Elizabeth	She will be down in a moment, I dare say. He then shut the door, and, coming up to her, claimed ...
Elizabeth	61	Elizabeth	Are you consulting your own feelings in the present case, or do you imagine that you are gratify...

Jane	62	Jane	On this point she was soon satisfied; and two or three little circumstances occurred ere they par...
Elizabeth	88	Elizabeth	But do you think she would be prevailed upon to go back with us? Change of scene might be of serv...
Checkpoint: step500000			
Elizabeth	0	Elizabeth	She was still very poorly, and [MASK] would not quit her at all, till late in the evening, when s...
Elizabeth	3	Elizabeth	I would go and see her if I could have the carriage. [MASK], feeling really anxious, was determi...
Elizabeth	10	Elizabeth	At any rate, she cannot grow many degrees worse, without authorising us to lock her up for the re...
Elizabeth	12	Elizabeth	[MASK] could safely say that it was a great happiness where that was the case, and with equal sin...
Elizabeth	75	Elizabeth	[MASK] read on: I have seen them both. They are not married, nor can I find there was any intent...
Elizabeth	76	Elizabeth	He has made me so happy, said she, one evening, by telling me that he was totally ignorant of ...
Elizabeth	77	Elizabeth	Indeed, replied [MASK], I am heartily sorry for him; but he has other feelings, which will pro...
Elizabeth	88	Elizabeth	But do you think she would be prevailed upon to go back with us? Change of scene might be of serv...
Checkpoint: step557000			