Remember to Forget: A Study on Verbatim Memorization of Literature in Large Language Models*

Xinhao Zhang[1,*], Olga Seminck[1,*] and Pascal Amsili[1]

[1]Lattice (UMR 8094, CNRS, ENS-PSL, Sorbonne Nouvelle), 1 rue Maurice Arnoux, 92120 Montrouge, France

Abstract

We examine the extent to which English and French literature is memorized by freely accessible LLMs, using a name cloze inference task (which focuses on the model's ability to recall proper names from a book). We replicate the key findings of previous research conducted with OpenAI models, concluding that, overall, the degree of memorization is low. Factors that tend to enhance memorization include the absence of copyrights, belonging to the Fantasy or Science Fiction genres, and the work's popularity on the Internet. Delving deeper into the experimental setup using the open source model Olmo and its freely available corpus Dolma, we conducted a study on the evolution of memorization during the LLM's training phase. Our findings suggest that excerpts of a book online can result in some level of memorization, even if the full text is not included in the training corpus. This observation leads us to conclude that the name cloze inference task is insufficient to definitively determine whether copyright violations have occurred during the training process of an LLM. Furthermore, we highlight certain limitations of the name cloze inference task, particularly the possibility that a model may recognize a book without memorizing its text verbatim. In a pilot experiment, we propose an alternative method that shows promise for producing more robust results.

Keywords

memorization, Large Language Models, membership inference attacks, literature, cloze task

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

*Corresponding author.

†These authors contributed equally.

zhangxinhao672@gmail.com (X. Zhang); olga.seminck@cnrs.fr (O. Seminck); Pascal.Amsili@ens.fr (P. Amsili)

https://github.com/XINHAO-ZHANG/ (X. Zhang);

https://www.lattice.cnrs.fr/membres/ingenieurs/olga-seminck/ (O. Seminck); https://lattice.cnrs.fr/amsili/ (P. Amsili)

0009-0003-7249-2091 (X. Zhang); 0000-0003-4617-5992 (O. Seminck); 0000-0002-5901-5050 (P. Amsili)

1. Introduction

The emergence of Large Language Models (LLMs) has advanced the field of Natural Language Processing (NLP) significantly. Successive models have consistently set new records on language understanding benchmarks [36, 35, 22]. Notably, LLMs can now tackle a broad range of tasks, allowing a single, general-purpose model to handle many NLP tasks. In the past, this required specialized models for each specific task. This shift has significantly increased the accessibility of NLP techniques, even for those without a specialized background. The ability to interact with LLMs through natural language, particularly via chat interfaces, has partially eliminated the need for programming knowledge.

These features have made LLMs ubiquitous, enabling their use for a wide range of purposes, including within the field of Digital Humanities, where they offer new perspectives. In addition to their ability to focus on specific tasks by learning from data curated by researchers [e.g. 16, 11], they also come equipped with pre-built knowledge and can be used even when there are no, or very few specific data at hand: the so-called zero-shot learning framework [e.g. 21, 5].

While the knowledge acquired during the training phase enables an LLM to function with few or no additional training data, this pre-training practice also presents several drawbacks and risks. One of the primary issues is that we lack a clear understanding of the specific knowledge these models possess, when of course this knowledge is crucial for accomplishing the tasks we give them.

The primary reason for this issue is that, for nearly all models, the specific data used for training remain unknown. When models are made available on platforms such as Hugging Face, users can typically access the model weights, but the training corpus itself is often not disclosed.

The second reason is that the actual learning process of such models is largely unknown, particularly regarding what determines whether certain data are remembered or forgotten. During training, billions of parameters are automatically adjusted within the model's neural network, and once this process is complete, it becomes impossible to interpret the activity of individual neurons. In this regard, these models are often referred to as "black boxes": the processes that generate a model's response to a user's task or question are virtually impossible to interpret. The main way to get an idea of a model's knowledge is to query it systematically and analyze its answers, but it still remains to be seen to what extent this allows us to get a full view of the knowledge. After all, even a slight change in the user's input can lead to significant variations in the results [13] and some models' outputs are not stable anyway (non-determinism).

Lacking a clear understanding of LLMs' knowledge presents a significant obstacle to their use in the field of Digital Humanities. We concur with Underwood [33] that a model's knowledge carries with it a certain world view and, consequently, a view of culture. When querying a model about literature, the texts included in its training corpus play a crucial role, as they fundamentally shape its understanding of the subject [12]. Questions regarding aesthetics, style, poetics, and so on will yield responses colored by the specific literature the model was trained on. Furthermore, it is essential to assess what a model retains from the books encountered during its training phase.

These questions are important not only in the context of literary research, but also for copyright compliance. If work covered by copyright is —unfortunately — in the training data, it is

important to be able to estimate to what extent it can be reproduced.

In this paper, we aim to address the extent to which literature is memorized by LLMs and the factors that contribute to this memorization. Additionally, we investigate whether it is possible to determine if work protected by copyright is in the training data of LLMs.

Our starting point is Chang, Cramer, Soni, and Bamman's study [8] who used a name cloze task to determine to what extent OpenAI's ChatGPT and GPT4 models are able to reproduce literary works verbatim (word for word). We applied the same method with freely accessible models, for English and French literature. In addition, we conducted a number of supplementary studies to gain a deeper understanding of the memorization process during training as well as the possible influence of the practice of prompting.

2. Related Work

Memorization in LLMs is generally defined as the verbatim reproduction of the training data [24, 3]. The phenomenon is typically associated with overfitting [7, 37]. It has been found that the following aspects can have a significant impact on memorization: data repetition in the training corpus, the number of model parameters (more parameters leading to a higher degree of memorization), and the number of tokens of context used to prompt the model [6].

Memorization is undesirable for various reasons. The first — and the most extensively studied by researchers — is that it includes privacy risks: generative models could disclose personal information (e.g. including URLs, phone numbers, and addresses) in their output if it has been memorized verbatim from the training data, making LLMs vulnerable to *training data extraction attacks* [6, 30, 3]. In the case of fiction, the privacy risk is less salient, but it is important that LLMs do not reproduce copyrighted material [15]. Furthermore, there are also risks of the memorization of literature from the public domain: as D'Souza and Mimno [9] stated: '*LLMs are poised to perpetuate the echoic nature of the literary canon within a new digital context*'. That is to say: the view of what is literature and what is not will be more and more influenced by how LLMs perceive it, because the number of applications of these models will only increase in the future, not only in the domain of literary studies, but in the entire culture sector where decisions about what should be commercialized are increasingly data driven [34].

Finally, in the context of literature, there is also the question of whether certain copyrighted works have been used to train LLMs. Memorization provides a lever to answer this question: if the model can be prompted to reproduce specific passages, it is an indication that the work has been used during training. Prompting a model to discover which data were present in the training set is called a *membership inference attack* [32]. Chang, Cramer, Soni, and Bamman [8] used this framework to study the verbatim memorization of literature by the LLMs of OpenAI: ChatGPT and GPT4. They found a high degree of memorization for some copyrighted works and an influence of the popularity of a book on the Internet with respect to the degree of memorization (popular books were better memorized), but the effect of memorization on downstream tasks remains equivocal. They expressed their concerns about the biases induced by memorization for studies in the field of cultural analytics where LLMs are used. They proposed the use of open models (with freely accessible training data) as a solution to the use of LLMs in the field of Digital Humanities.

In the remainder of this paper, we present the name cloze task proposed in [8], that we used and adapted for English and French with a variety of freely available models (section 3.1); we report and discuss the results that we obtained in section 3.3, along with several analyses of the behaviour of the models depending on the copyright status, sub-genre, and popularity of the works chosen to probe the models. We also present further studies that we ran to get a better understanding of the learning, memorization and recalling processes. These are presented in sections 3.4 and 4.

3. Name cloze task

3.1. Task

To assess the memorization of literary data by language models, Chang, Cramer, Soni, and Bamman [8] formulated a *membership inference attack* task, which they call *name cloze inference*, where models have to predict a proper name missing from a text passage. Unlike other completion tasks focusing on predicting named entities [17, 27], the text passages used by Chang, Cramer, Soni, and Bamman [8] contain no other named entities than the target name. Therefore, this type of task tests the models' ability to 'remember' very specific information from the training data. By way of comparison, human performance on this task was assessed at 0% by Chang, Cramer, Soni, and Bamman [8]: the contexts were not informative enough for humans to guess the target names.

The experiments presented in this section used the protocol of Chang, Cramer, Soni, and Bamman [8]. We used the prompt presented in Figure 1 that displays two examples (that did not vary across items) followed by the target item.

3.2. Data

The items we used for the task were taken from Chang, Cramer, Soni, and Bamman [8] for the English experiment (3.2.1), and we used a similar method to construct the items for the French experiment (3.2.2).

3.2.1. English

Chang, Cramer, Soni, and Bamman [8] created an item set by running the BookNLP[1] pipeline [1] on the literary corpus presented in Table 1 to extract passages with a proper name of the type *character* and no other named entities. They then randomly sampled 100 passages per book. Books with fewer than 100 passages were excluded from the experiment. In total, there were 57,100 items.[2] Two examples are given below:

(1) a. 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.

[1] https://github.com/booknlp/booknlp

[2] Items generated from these books can be found in a github repository: https://github.com/bamman-group/gpt4-books/tree/main/data/model_output/chatgpt_results

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 1: Prompt for Name Cloze Inference. The prompt is almost identical to that of Chang, Cramer, Soni, and Bamman [8], the difference is that we added the sentences *'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:'*. The examples are identical. We made these decisions based on preliminary tests performed on Mixtral8x7B [20]. This prompt was used for English and French. After some preliminary testing, we decided not to translate for French, as this seemed to lead to results of lower quality.

b. 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.

Items from the book *Pride and Prejudice*

3.2.2. French

The French item set was selected from the Chapitres corpus [23], which includes about 3,000 digitized books in French. Thanks to the fr-BookNLP pipeline [26], we were able to easily extract passages from books and produce items in the same manner as Chang, Cramer, Soni, and Bamman [8]. Each of the items contains exactly one proper name of a character (named entity of type PERSON) as a single token (see Example (2)).

(2) a. Le campagnard, à ces mots, lâcha l'étui qu'il tournait entre ses doigts. Une saccade

Table 1
Number of books selected by collection and genre for English.

Genre	#Books
LitBank collection [2]	91
Novels nominated for the Pulitzer Prize	90
Bestsellers listed by *NY Times* and *Publishers Weekly*	95
The Black Book Interactive Project & the Black Caucus American Library Association [3]	101
Global Anglophone fiction (outside the U.S. and U.K.)	95
Science fiction, fantasy, horror, mystery, crime, romance and spy novels	99
Total	**571**

de ses épaules fit craquer le dossier de la chaise. Son chapeau tomba.– Je m'en doutais, dit [MASK] en appliquant son doigt sur la veine.

b. En passant auprès des portes, la robe d'[MASK], par le bas, s'ériflait au pantalon ; leurs jambes entraient l'une dans l'autre ; il baissait ses regards vers elle, elle levait les siens vers lui ; une torpeur la prenait, elle s'arrêta. Items from the book *Madame Bovary*

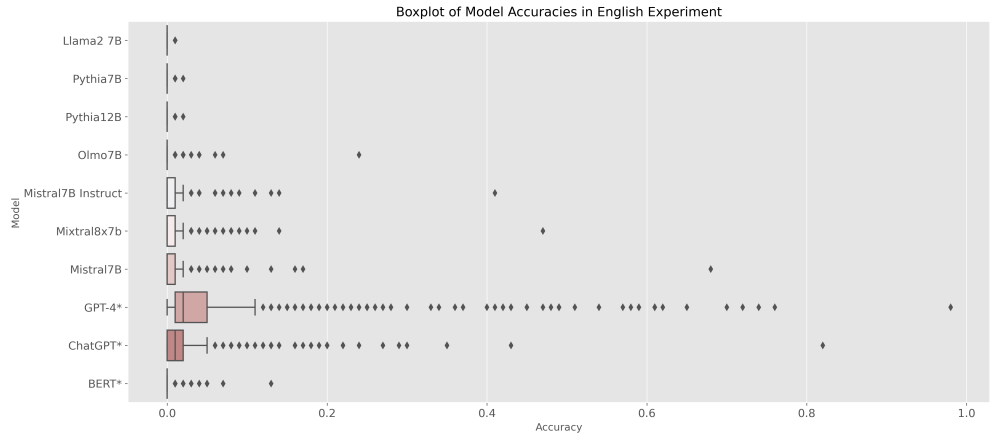
After excluding books with fewer than 100 generated elements, 2,459 books remained. However, limiting the number of books is still necessary in order to avoid an excessive experiment runtime. We selected 575 French books by balancing per genre, as shown in Table 2. For all books, we also carried out a random selection of 100 items each.

Table 2
Breakdown by genre of the 575 books that were selected from the Chapitres corpus [23] to build the French item set.

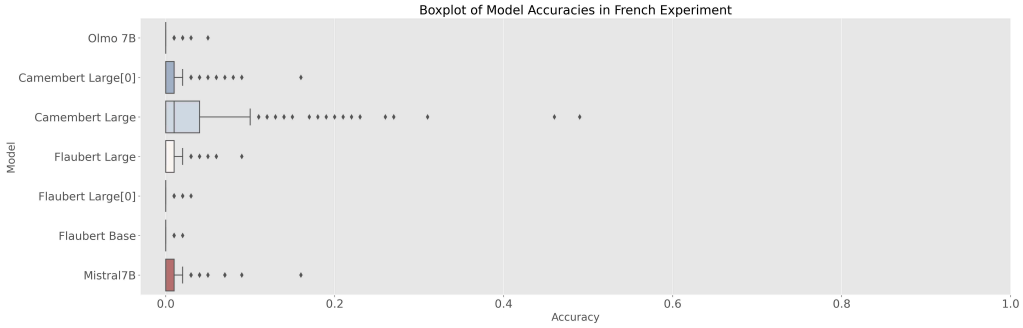
Genre	#Books
Thriller	111
Adventure novels	109
Children's literature	99
Historical fiction	96
Cycles and series	79
Short stories	78
Total	**575**

3.3. Replication

In this section, we report on the replication of Chang, Cramer, Soni, and Bamman's name cloze inference task using freely accessible models. The data we used are described in the previous subsection.



(a) English. The accuracies marked with an asterisk (*) are results reported by Chang, Cramer, Soni, and Bamman [8].



(b) French. For CamemBERT or FlauBERT, [0] means that we only counted a *hit* if the highest ranking answer was the correct proper name. For the other versions, we considered that there was a hit if the correct answer was among the top 5 highest ranking answers.

Figure 2: Box-plots of the scores of various models in English and French on the name-cloze inference task.

3.3.1. Replication with open models

English: We tested MistralAI (Mistral7B, Mistral7B-Instruct and Mixtral8x7B) [19, 20], Olmo7B [14], Pythia (7B et 12B) [4] and Llama2 7B [31], in order to compare the performance of all these models. For the ChatGPT, GPT-4 and BERT [10] models, the scores were taken directly from the data of Chang, Cramer, Soni, and Bamman [8]. The performance of each model on the task is plotted in Figure 2a.

First, we observe that, with an average accuracy of 6.81%, GPT-4 clearly stands out as the best-performing model, followed by ChatGPT (GPT 3.5 turbo) with an average score of 2.51%. The Mistral7x8b, Mistral7B and Mistral7B-Instruct models show scores only just under 1%. The other models (Olmo 7B, BERT, Pythia12B, Pythia7B and Llama2 7B) show lower accuracies, ranging from 0.27% to 0.01%.

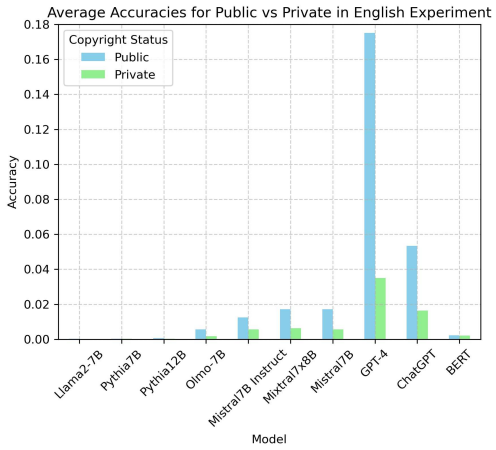
Interestingly, the vast majority of books score (close to) 0%. The outliers are relatively few in number, and it is probably only for these that we can speak of memorization. Intriguingly, for almost all models (except BERT), the text *Alice's Adventures in Wonderland* obtains the highest scores, probably due to its notoriety and high frequency in the training corpus.

French: We decided not to test all the models we tested for English. As running these models is time and resource consuming (about one night per model and even a whole week for Mixtral8x7B) on our server with one GPU, we decided to exclude Mixtral8x7B because of its consumption and unexceptional level of memorization and Mistral7B-Instruct, Llama2 and all the versions of Pythia because of very low degrees of memorization. To replace BERT for English, we introduced comparable models specialized for French: CamemBERT [25] and FlauBERT [22]. The scores of these models can be found in Figure 2b.

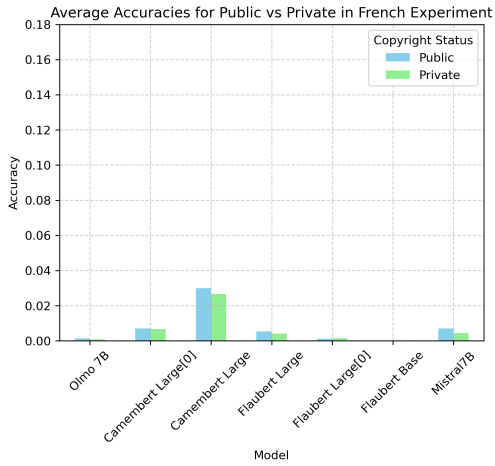
Remarkably, for French, the language-specialized model CamemBERT performed by far the best, and in contrast to English where the BERT model was one of the lowest scoring compared to latest generation LLMs, the BERT-architecture models for French performed similarly to Mistral7B and better than Olmo7B.

3.3.2. Analysis of copyright status

Figure 3a shows the accuracy of the models according to copyright status. A general trend can be observed: all models scored higher for public works for English and French, even though the difference is smaller for French. This result confirms our hypothesis that the models are mainly trained on public domain books, and replicates the findings from Chang, Cramer, Soni, and Bamman [8].



(a) Average accuracy of books from the public domain (public) and under copyright (private) for English.



(b) Average accuracy of books from the public domain (public) and under copyright (private) for French.

Figure 3: Comparative accuracy of books based on copyright status in English and French.

3.3.3. Analysis of the sub-genres of books

We have already noted that freely accessible LLMs can predict certain elements from books, regardless of their copyright status. Table 3 explores this capability by detailing the performances by specific genres of the sub-corpus in English.

Apart from a significant difference in accuracy scores, the trends observed on the English items are similar to those of Chang, Cramer, Soni, and Bamman [8]. The tested models seem to have the best knowledge of science fiction and fantasy works and public domain texts. However, they are less familiar with Global Anglophone fiction and works from black authors. For French, we observe that CamemBERT, Flaubert and Mistral7B obtain the highest score on children's literature and Olmo7B on historical novels (see Table 4).

Table 3
Name cloze average accuracy regarding sub-genres of books in the English experiment. Numbers in bold are the highest scores per column.

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 4
Name cloze average accuracy regarding sub-genres of books in the French experiment. Numbers in bold are the highest scores per column.

Literary genre	Olmo-7B	Camembert Large[0]	Camembert Large	Flaubert Large	Flaubert Large[0]	Flaubert Base	Mistral7B
Cycle and series	0.0008	0.0082	0.0272	0.0052	0.0016	0.0003	0.0072
Children's literature	0.0012	**0.0099**	**0.0481**	**0.0079**	0.0011	0.0002	**0.0093**
Short stories	0.0012	0.0086	0.0296	0.0059	0.0014	**0.0005**	0.0086
Thriller	0.0005	0.0023	0.0136	0.0018	0.0005	0.0000	0.0025
Adventure novels	0.0011	0.0050	0.0191	0.0041	0.0015	0.0003	0.0057
Historical fiction	**0.0025**	0.0085	0.0372	0.0058	**0.0017**	0.0002	0.0054

On the one hand, it certainly makes sense that the models perform better on public domain texts, due to the regulations on the use of free works. On the other hand, the specificity of the science fiction and *fantasy* genres seems to facilitate the models' prediction. By closely examining items from the *'Science-Fiction/Fantasy'* genre, we found words that are not named entities but that are still very indicative of the book, such as for instance 'Quidditch', 'Witchcraft', or 'Muggles' in items from *Harry Potter*.

3.3.4. Analysis of book popularity on the web

According to Chang, Cramer, Soni, and Bamman [8], a book's popularity should be defined by its presence in many academic libraries, its frequency in large-scale training datasets (such as Books3, part of The Pile), its citations in non-indexed academic journals, and its appearance on the public web (both in excerpts and full text). In line with Chang, Cramer, Soni, and Bamman [8], we checked whether there was a relationship between the popularity of a book online and the degree of memorization of models for the English items. We used the number of hits from Bing, Google and the C4 corpus directly from their data and calculated a Spearman's correlation with the accuracy scores of the freely accessible models that we tested.

Most open language models showed a positive correlation between prediction performance and book popularity on the web (see Table 5). This experiment therefore reinforces the hypothesis that web prevalence is correlated with performance on the name-cloze inference task. However, the models that performed poorly (i.e. those that failed to give the right prediction for most books) do not show a high correlation with any engine/corpus. It is for this reason that we decided not to repeat this experiment for French: as generative LLMs perform poorly on the French dataset, we did not expect high correlations between the accuracy on the French items and the popularity of a work online.

Table 5
Spearman's correlation between model accuracy and the online popularity of books from the English data set.

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

3.4. Evolution of memorization during training

Since a high degree of memorization was found for some books and some models, and since the popularity of a work online is correlated with the performance of the models, it seems natural to wonder whether memorizing a book requires access to the full text, or if it can also take place via excerpts from websites. In this section, we therefore present a new series of experiments, in which we monitored the memorization of books during the pre-training process of an LLM. Inspired by Biderman, Schoelkopf, Anthony, Bradley, O'Brien, Hallahan, Khan, Purohit, Prashanth, Raff, et al. [4] and Biderman, Prashanth, Sutawika, Schoelkopf, Anthony, Purohit, and Raff [3], we studied the emerging pattern of memorization as a function of the

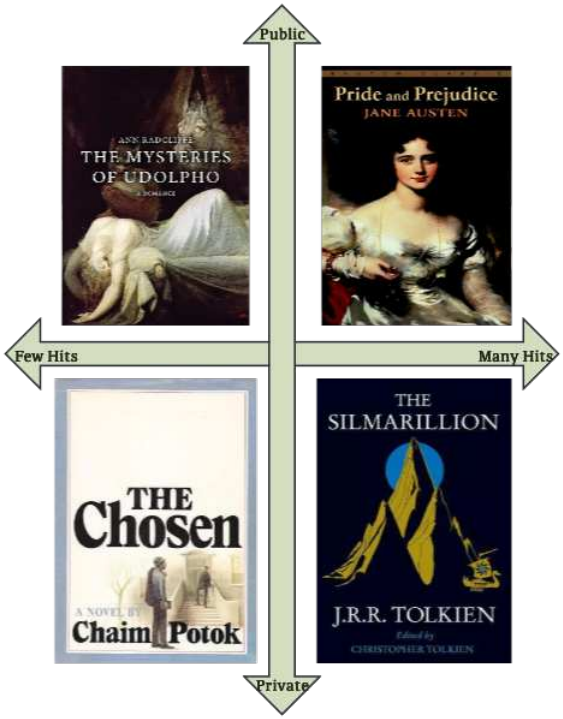


Figure 4: Four books selected based on two criteria: copyright status and the popularity of the works online as measured by Chang, Cramer, Soni, and Bamman [8].

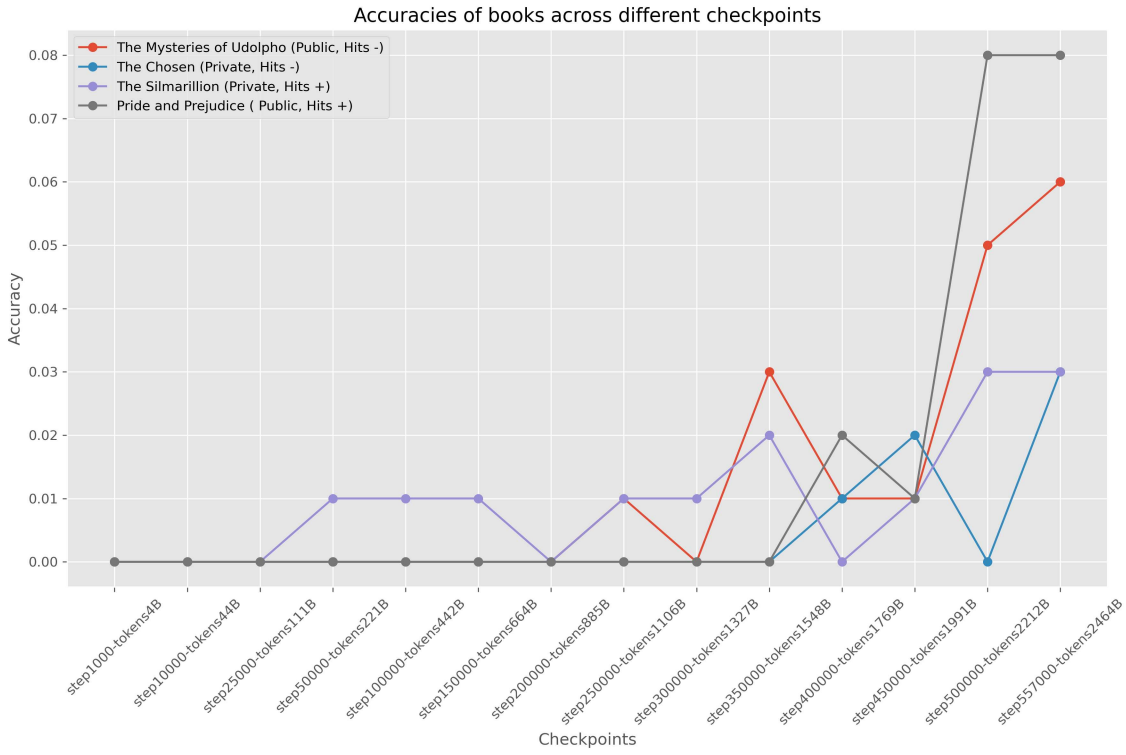


Figure 5: Evolution of accuracy scores across different checkpoints

book's popularity online and whether it is in the public domain or under copyright.

For this experiment, we used the OLMo7B model [14] as it has been trained on fully public data, the Dolma corpus [29] and provides numerous checkpoints (states of the models during the pre-training phase).

It is beyond our computational resources to run experiments for all 571 books on OLMo's more than 500 checkpoints. (As many OLMo models would have to be downloaded as there are *checkpoints*; i.e. more than 500, and the experiment would therefore take 500 times longer than the initial experiment with this model.) That is why, in our study, we focused on fourteen *checkpoints* — chosen at regular intervals — and four particularly representative books, selected according to two dimensions, as illustrated in Figure 4: copyright status (public or private), and their popularity (few hits or many hits). These works are respectively *The Mysteries of Udolpho*, *Pride and Prejudice*, *The Chosen* and *The Silmarillion*.

Figure 5 shows the evolution of memorization during the training of OLMo. For the works in the public domain (*The Mysteries of Udolpho* and *Pride and Prejudice*) there is a noticeable increase in accuracy towards the end of training, particularly between steps 450,000 and 557,000. It can reasonably be suggested that at this stage of training, the model is seeing the full texts of free works, such as those available in the most reputable projects such as Project Gutenberg. This hypothesis is reinforced by the observation that in the Dolma corpus [29] corpora representing literature are placed at the end.[4]

In contrast, for the copyrighted works, *The Chosen* and *The Silmarillion*, their performance evolved continuously and steadily throughout the training period, without showing such a sharp and sudden increase. For example, right from the start of the pre-training phase, from step 50,000 onwards, the OLMo model successfully predicted a masked proper noun in *The Silmarillion* items. For these works, the accuracy fluctuated slightly but remained relatively stable throughout the training phase, right up to the end, although there were some additional good predictions. This could support the hypothesis that excerpts or quotations from this book are scattered throughout various sub-corpora and distributed throughout the pre-training phase. Furthermore, it is clear that the influence of web popularity, measured by the number of 'hits', also plays an important role in evolution, especially for copyrighted works. This is particularly true for *The Silmarillion*, whose popularity on the web is associated with more pronounced fluctuations in predictive scores.

3.5. Discussion

The experiments in this section on the name cloze tasks first show that most models do not feature a high degree of memorization in general. However, for some particular works the degree of memorization can be very high. Despite the fact that average scores for ChatGPT and GPT4.0 were higher, our data show the same distribution as Chang, Cramer, Soni, and Bamman [8]'s, for English and for French. Interestingly, our experiments suggest that the number of parameters is not a determining factor for memorization: heavier models from the same series do not show an enhancement in accuracy on the task (e.g. Pythia13B with respect to Pythia7B and Mixtral8x7B versus Mistral7B). For French, it is noteworthy that the BERT-type models were the highest performing models, in contrast to English. Our hypothesis is

[4]Unfortunately, we could not find a map explaining which checkpoint corresponded exactly to which part of Dolma.

that there might be a higher overlap between the pre-training corpus of CamemBERT and FlauBERT and the French items we constructed than there is between the items for English and the pre-training corpus of BERT. We also think that the amount of training data in French, which is smaller than the amount of English training data, must play an important role.

In our experiments, we also replicated Chang's findings that public domain books were better remembered by LLMs than copyrighted books; we found this for both English and French. We also replicated the relationship between the online popularity of books and scores on the name cloze task, although this relationship was not strong for books for which LLMs showed low levels of memorization anyway. Also, for the English items, we replicated the finding that books from the genre of science fiction and fantasy were better memorized than those from other genres.

However, during the replication with open models we ran into various problems with the protocol of the name cloze task. In section 3.3.3, we already identified the problem of words that are not named entities, but are very specific to a particular book (*e.g. Muggles* in *Harry Potter*). Moreover, during our experiments, we also saw that some items do contain named entities that are not detected by BookNLP (for example, 'Hogwarts' and 'Voldemort' in *Harry Potter*). Also, style is sometimes very recognizable, for example — to stay with the example of *Harry Potter* — the way the character Hagrid speaks (see example (3)).

(3) "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]?"

This suggests that it is possible that instead of recognizing verbatim a sentence from the training data, a model recognizes a book based on specific vocabulary, unfiltered named entities and style, and guesses the name of the main character. This strategy would lead to a high performance, as we checked for the English items that the main character was the correct answer 29.48% of the time, which is much higher than the performance of any LLM on the name cloze inference task.

Another concern that we have about the name cloze task is the exclusive focus on proper names. A proper name might not be the most representative morpho-syntactic category for all words. Indeed, Pang, Ye, Wang, Yu, Wong, Shi, and Tu [28] found in a morpho-syntactic analysis carried out in the context of LLMs that proper nouns are systematically given higher attention weights than common nouns or other word types.

Finally, we also question whether prompting is the most ideal way to access the memory of LLMs. We wonder if the lower scores we found for open models with respect to Chang, Cramer, Soni, and Bamman [8]'s findings on OpenAI models can be explained by a better chat-module of the latter, i.e. : it could be the case that memorization seems lower than it is for open models because memory cannot be accessed conveniently by prompting (the comprehension of instructions might be higher for the OpenAI models).

4. Further analysis

These concerns with the name cloze task led us to design two new experiments: the first aims at checking whether the prompting framework is suited to querying open LMMs (section 4.1) and the second proposes an alternative protocol to the name cloze inference task (section 4.2).

4.1. Evaluating the appropriateness of prompting for the name cloze task

In this section, we present a fine-tuning experiment of the Mistral7B model [19] to assess whether prompting influences model performance on the name cloze task. The idea is the following: we seek to enhance the task comprehension by fine-tuning the LLMs on English items from books from the public domain. These books are certainly in the training data because they are widely available for example in the Project Gutenberg[5] or on Wikibooks[6]. Our hypothesis is that if books have been memorized, the fine-tuning helps the model to learn how to access the information from its memory.

An example of an item from the fine-tuning training data is shown below:

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

Regarding the fine-tuning method, we employed Lora [18], a model quantization technique available in the Python library *peft*[7]. The fine-tuned model has been integrated and is accessible on our Hugging Face account's site[8], where it is presented with the results of the fine-tuning experiment.

The evolution of the loss value is shown in Figure 6. It can be observed that this value decreases significantly only during the initial steps. The average *accuracy* score of the Mistral7B model without fine-tuning is 0.00830, while the fine-tuned version achieves a score of 0.00893, so fine-tuning did not yield substantial gains on the task's performance. We conclude that the fact that open models fail at the name cloze inference task cannot be explained by a misunderstanding of the prompt.

4.2. Pilot experiment: study memorization with n-grams

Memorization of proper names may not be representative for other part-of-speech categories. Therefore, we conducted a pilot experiment to evaluate the use of an alternative method to the name cloze inference task. The idea is very simple: we ask an LLM to complete a passage

[5] https://www.gutenberg.org
[6] https://www.wikibooks.org
[7] https://pypi.org/project/peft/
[8] https://huggingface.co/LivevreXH/mistral_finetuned_items_livres/tree/main

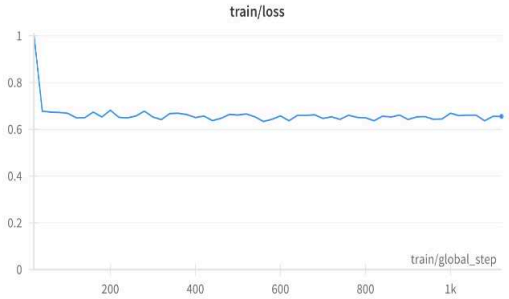


Figure 6: Evolution of training loss during fine-tuning. After a first gain in performance, the model quickly stagnates.

extracted from a book and count the overlap of the first ten tokens it produces with the real text in the book. For this pilot, we took the four books presented in Figure 4 and used the corresponding items from Chang, Cramer, Soni, and Bamman [8] in the following manner: first we replaced the [MASK]-token with the proper name, and then we took the first ten tokens to be presented in the prompt and the following 10 tokens as a gold answer. Our prompt is provided in Figure 7. To compare this method to the name cloze inference task, we decided to test ChatGPT and study the correlation between the scores on the two tasks. The results can be found in Figure 8.

As a sanity check, we also established a baseline score for the n-gram method. A young novelist, Jingyi, provided us with an unpublished draft of her next novel, written in Chinese. We translated this text into English using the DeepL translation tool[9]. From the translated manuscript, we selected 100 random excerpts. We submitted this manuscript to the same prediction task. The memorization score was very low: 0.005. In comparison, the lowest scoring novel from Figure 8 obtained a score of 0.038, more than seven times as high.

The number of books tested in this framework remains low and therefore the performance of the pilot should be interpreted with caution. Still, we want to put forward a first evaluation of the n-gram method as opposed to the name cloze inference. A first observation is that both tasks show a substantial level of correlation (0.77) but that the values of the scores for the n-gram task are more fine-grained than those of the name cloze task. Indeed, whereas for the name cloze task we have 100 items per book, for the n-gram task we have 100 x 10 tokens to evaluate which can help to make a better distinction amongst the lower scoring works. The baseline of the unseen manuscript shows that there still is some distinction to make between very low degrees of memorization and no memorization at all.[10] Furthermore, our results suggest that the n-gram method could help against the sensitivity of the name cloze task to recognizing a style, or specific word from a fictional universe and guessing a random character from a work without true memorization of the exact passage. Looking at "The Silmarillion" in Figure 8, we see that its n-gram score is lower than would be expected by looking at the name cloze inference score. Inspecting Chang, Cramer, Soni, and Bamman's items for this book more

[9] https://www.deepl.com/fr/translator

[10] Admittedly, the translation of a Chinese novel by DeepL might not be the most representative literature and this experiment should be repeated using an unpublished draft of a native speaker writer.

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 7: Prompt of the n-gram pilot experiment.

closely, we observe that there are important differences in the choice of answers of ChatGPT. For example: 8 items should receive the answer 'Melkor' but ChatGPT never put forward this name, whereas it predicts 'Aragorn' 4 times even though this is never the correct answer. This leads us suspect that the name cloze task is sensitive to the short cut of guessing a character from a book rather than retrieving the correct name from its memory.

5. Conclusion

The memorization of English and French literature is low on average in freely accessible LLMs, while a small number of fictional works seem to undergo an extreme degree of memorization. Memorization is favored by the presence of quotes and excerpts of the books on the Internet, which makes it impossible to say if a high score for memorization means that the full text of the novel was actually used to train an LLM, except if the training corpus has been released, which is only the case for a very small number of LLMs.

For our research, we used the name cloze inference task, in which an LLM must guess a proper name from a sentence without the presence of any other named entities. Using this method, it occurred to us that it has some undesirable effects that were initially unforeseen. The first is that the method is sensitive to errors. As items are automatically filtered for named entities, not all named entities are removed from the context and could be used by the LLM to

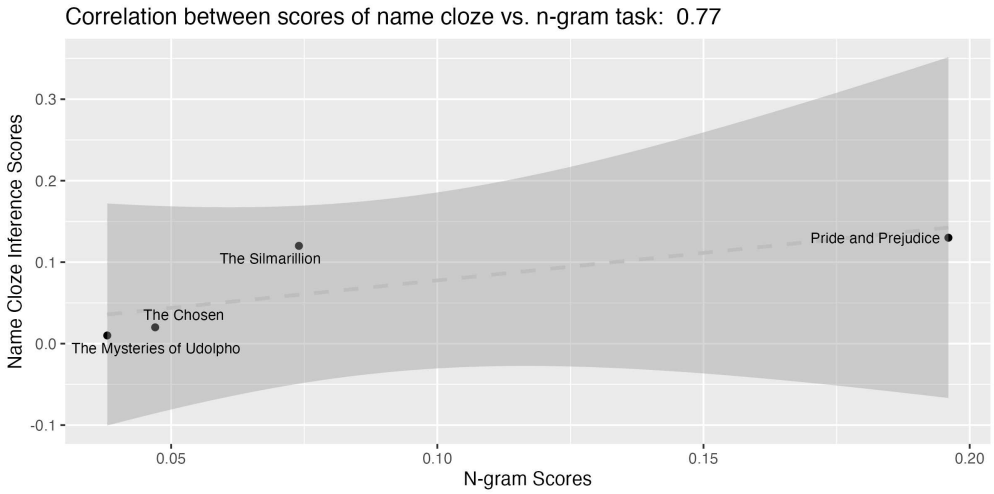


Figure 8: The correlation between the scores on the name cloze inference task and the n-gram task for the ChatGPT model on the four selected books from Figure 4.

guess the name of a character from the book without there being real verbatim memorization. The same can happen because of a recognizable style and typical words (such as in science fiction novels). Given the fact that the memorization score of LLMs is low, this noise cannot be ignored. When testing a very simple alternative method that counts n-gram overlap when the model is prompted to continue a passage from a novel, our pilot experiment showed that this method has the potential to be more robust than the name cloze inference task.

In future work, we aim to explore not only verbatim memorization, but also memorization of plots and stories. Ultimately, coming back to the introduction in which we argued that LLMs give a biased point of view on culture and literature, we would like to not only measure the spread and memorization of exact texts, but also of ideas and more abstract patterns present in literature.

6. Availability of Resources and Code

All the experimental items and programming code for our experiments can be found on the following GitHub page: https://github.com/XINHAO-ZHANG/books-memorization.

Acknowledgments

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the ”Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute, Thierry Poibeau’s Chair).

References

[1] D. Bamman. *BookNLP*. 2021. URL: https://github.com/booknlp/booknlp.

[2] D. Bamman, O. Lewke, and A. Mansoor. "An Annotated Dataset of Coreference in English Literature". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis. Marseille, France: European Language Resources Association, 2020, pp. 44–54. URL: https://aclanthology.org/2020.lrec-1.6.

[3] S. Biderman, U. Prashanth, L. Sutawika, H. Schoelkopf, Q. Anthony, S. Purohit, and E. Raff. "Emergent and predictable memorization in large language models". In: *Advances in Neural Information Processing Systems* 36 (2024).

[4] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, et al. "Pythia: A suite for analyzing large language models across training and scaling". In: *International Conference on Machine Learning*. Pmlr. 2023, pp. 2397–2430.

[5] J. Borst, J. Klähn, and M. Burghardt. "Death of the Dictionary?–The Rise of Zero-Shot Sentiment Classification". In: *CHR 2023: Computational Humanities Research Conference*. 2023.

[6] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang. "Quantifying Memorization Across Neural Language Models". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=TatRHT%5C%5F1cK.

[7] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel. "Extracting Training Data from Large Language Models". In: *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 2021, pp. 2633–2650. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

[8] K. Chang, M. Cramer, S. Soni, and D. Bamman. "Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, 2023, pp. 7312–7327. DOI: 10.18653/v1/2023.emnlp-main.453.

[9] L. D'Souza and D. Mimno. "The Chatbot and the Canon: Poetry Memorization in LLMs". In: *CHR 2023: Computational Humanities Research Conference*. 2023.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *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)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[11] G. G. Garcia and C. Weilbach. “If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History”. In: *CHR 2023: Computational Humanities Research Conference*. 2023.

[12] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. “Datasheets for datasets”. In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[13] H. Gonen, S. Iyer, T. Blevins, N. Smith, and L. Zettlemoyer. “Demystifying Prompts in Language Models via Perplexity Estimation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, 2023, pp. 10136–10148. DOI: 10.18653/v1/2023.findings-emnlp.679.

[14] D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. H. Jha, H. Ivison, I. Magnusson, Y. Wang, et al. “Olmo: Accelerating the science of language models”. In: *arXiv preprint arXiv:2402.00838* (2024).

[15] P. Henderson, X. Li, D. Jurafsky, T. Hashimoto, M. A. Lemley, and P. Liang. “Foundation Models and Fair Use”. In: *Journal of Machine Learning Research* 24.400 (2023), pp. 1–79. URL: http://jmlr.org/papers/v24/23-0569.html.

[16] R. M. Hicke and D. Mimno. “T5 meets Tybalt: Author Attribution in Early Modern English Drama Using Large Language Models”. In: *CHR 2023: Computational Humanities Research Conference*. 2023.

[17] F. Hill, R. Reichart, and A. Korhonen. “SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation”. In: *Computational Linguistics* 41.4 (2015), pp. 665–695. DOI: 10.1162/COLI_a_00237. URL: https://aclanthology.org/J15-4004.

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL].

[19] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. “Mistral 7B”. In: *arXiv preprint arXiv:2310.06825* (2023).

[20] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al. “Mixtral of experts”. In: *arXiv preprint arXiv:2401.04088* (2024).

[21] P. Kaganovich, O. Münz-Manor, and E. Ezra-Tsur. “Style Transfer of Modern Hebrew Literature Using Text Simplification and Generative Language Modeling”. In: *CHR 2023: Computational Humanities Research Conference*. 2023.

[22] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab. “FlauBERT: Unsupervised Language Model Pre-training for French”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis. Marseille, France: European Language Resources Association, 2020, pp. 2479–2490. URL: https://aclanthology.org/2020.lrec-1.302.

[23] A. Leblond. *Corpus Chapitres*. Version v1.0.0. 2022. DOI: 10.5281/zenodo.7446728.

[24] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini. "Deduplicating Training Data Makes Language Models Better". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by S. Muresan, P. Nakov, and A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 8424–8445. DOI: 10.18653/v1/2022.acl-long.577.

[25] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Online: Association for Computational Linguistics, 2020, pp. 7203–7219. DOI: 10.18653/v1/2020.acl-main.645.

[26] F. Mélanie-Becquet, J. Barré, O. Seminck, C. Plancq, M. Naguib, M. Pastor, and T. Poibeau. *BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature*. Tech. rep. 1. Darmstadt, 2024, 34 Seiten. DOI: https://doi.org/10.26083/tuprints-00027396.

[27] T. Onishi, H. Wang, M. Bansal, K. Gimpel, and D. McAllester. "Who did What: A Large-Scale Person-Centered Cloze Dataset". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by J. Su, K. Duh, and X. Carreras. Austin, Texas: Association for Computational Linguistics, 2016, pp. 2230–2235. DOI: 10.18653/v1/D16-1241.

[28] J. Pang, F. Ye, L. Wang, D. Yu, D. F. Wong, S. Shi, and Z. Tu. *Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models*. 2024. URL: http://arxiv.org/abs/2401.08350.

[29] L. Soldaini, R. Kinney, A. Bhagia, D. Schwenk, D. Atkinson, R. Authur, B. Bogin, K. Chandu, J. Dumas, Y. Elazar, et al. "Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research". In: *arXiv preprint arXiv:2402.00159* (2024).

[30] R. Staab, M. Vero, M. Balunovic, and M. Vechev. "Beyond Memorization: Violating Privacy via Inference with Large Language Models". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=kmn0BhQk7p.

[31] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[32] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei. "Towards demystifying membership inference attacks". In: *arXiv preprint arXiv:1807.09173* (2018).

[33] T. Underwood. *Mapping the latent spaces of culture*. Essay prepared for a roundtable. 2021.

[34] M. Walsh. *Where is all the book data?* Online essay. 2022. URL: https://www.publicbooks.org/where-is-all-the-book-data/.

[35] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. "SuperGLUE: a stickier benchmark for general-purpose language understanding systems". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[36] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by T. Linzen, G. Chrupała, and A. Alishahi. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446.

[37] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. "Understanding deep learning (still) requires rethinking generalization". In: *Communications of the ACM* 64.3 (2021), pp. 107–115.