Evaluating LLMs for Quotation Attribution in Literary Texts: A Case Study of LLaMa3

Gaspard Michel[†*]
gmichel@deezer.com

Elena V. Epure[†]
eepure@deezer.com

Romain Hennequin[†]
rhennequin@deezer.com

Christophe Cerisara[*]
christophe.cerisara@loria.fr

[†] Deezer Research, Paris, France
[*] Loria, Nancy, France

Abstract

Large Language Models (LLMs) have shown promising results in a variety of literary tasks, often using complex memorized details of narration and fictional characters. In this work, we evaluate the ability of Llama-3 at attributing utterances of direct-speech to their speaker in novels. The LLM shows impressive results on a corpus of 28 novels, surpassing published results with ChatGPT and encoder-based baselines by a large margin. We then validate these results by assessing the impact of book memorization and annotation contamination. We found that these types of memorization do not explain the large performance gain, making Llama-3 the new state-of-the-art for quotation attribution in English literature. We release publicly our code and data[1].

1 Introduction

Quotation attribution, or the automated attribution of utterances to fictional characters, is of crucial importance for character analysis in digital humanities (Elson et al., 2010; Muzny et al., 2017a; Labatut and Bost, 2019; Sims and Bamman, 2020). However, quotation attribution remains a challenging task, and recent approaches still struggle to find methods that generalize across writing styles. A few works have explored the use of LLMs for quotation attribution in novels, by extracting conversations directly with ChatGPT (Zhao et al., 2024) or by asking ChatGPT to attribute a single quote given its surrounding context (Su et al., 2023). Yet, these works do not propose a systematic evaluation of LLMs for quotation attribution in literary works.

Another significant evaluation drawback in assessing LLMs is the lack of analysis regarding book memorization and annotation contamination, which can hinder their generalization abilities. Book memorization occurs when an LLM is able to generate specific passages of texts in a novel, and is correlated with its frequency in pretraining data (Carlini et al., 2023). In contrast, data contamination arises when an LLM has memorized evaluation data, enabling it to produce labels without reasoning (Magar and Schwartz, 2022). To avoid confusion, we refer to data contamination as annotation contamination. Addressing both issues is essential when evaluating LLMs on literary tasks, as they can significantly impact the understanding of its performance on downstream tasks.

In this work, we start by evaluating the performance of Llama-3 8b on the Project Dialogism Novel Corpus (PDNC) (Vishnubhotla et al., 2022), a corpus of 28 English novels. We selected Llama-3 8b due to its popularity, its impressive performance on various tasks (Dubey et al., 2024), and because its pretraining corpus only includes data up to March 2023, which makes the second release of PDNC annotations not included in the pretraining data. We carefully designed prompts with Chain-of-Thought reasoning (Wei et al., 2022), and use the larger context size of LLMs to directly attribute all quotes in a given chapter. Our results indicate that this method improves attribution accuracy compared to predicting a single quote in a contextual passage. We next conduct an evaluation of book memorization and annotation contamination to determine whether Llama-3's success stems from its reasoning abilities or its capacity to memorize passages and annotations.

We found that our Llama-3 based approach demonstrates remarkable performance, improving attribution accuracy by 12 points against state-of-the-art systems on the first 22 novels on PDNC and by 9 points on the remaining novels. Besides, we could not find signs of annotation contamination on the first 22 PDNC novels, and we show that although memorization impacts speaker predictions on a subset of quotes, a majority of successful predictions can be attributed to the reasoning ability of

[1] https://github.com/deezer/llms_quotation_attribution

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 742–755
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Llama-3. We validate this finding by evaluating the LLM on a recently published novel not included in its pretraining data, where our approach performs on-par with the current state-of-the-art system, BookNLP+ (Vishnubhotla et al., 2023; Michel et al., 2024). Besides, we found that our approach combined with the larger Llama-3 70b reaches an almost perfect accuracy. To sum up, our contributions are:

1. We evaluate Llama-3 zero-shot performance on PDNC, comparing it to strong systems and show a major accuracy improvement on PDNC novels, establishing a new state-of-the-art for quotation attribution accuracy on English literature.

2. We introduce a novel measure of book memorization, *Corrupted-Speaker-Guessing*, that classifies a successful quote attribution into either a reasoning or memorization prediction. We propose this new measure as other metrics (Chang et al., 2023) failed to detect memorization of canonical literature when used with Llama-3 8b. We validate our measure following a similar evaluation protocol as Chang et al. (2023).

3. We thoroughly evaluate the impact of book memorization and annotation contamination on the downstream task, showing that these memorization types are not the principal factors of Llama-3 quotation attribution accuracy.

"As soon as ever Mr. Bingley comes, my dear," said Mrs. Bennet, "you will wait on him of course."

"No, no. You forced me into visiting him last year, and promised if I went to see him, he should marry one of my daughters..."

His wife represented to him how absolutely necessary such an attention would be from all the neighbouring gentlemen, on his returning to Netherfield.

"'Tis an etiquette I despise," said he.

Figure 1: Excerpt of *Pride and Prejudice* by Jane Austen (1813). Quotations are colored by quote type: explict, implicit and anaphoric. Speaker information given by the narrator are underlined. Figure taken from Michel et al. (2024).

2 Related Work

LLMs for literary tasks Large Language Models (LLMs) have shown promising results in a variety of literary tasks related to Narrative Understanding (Xu et al., 2023; Underwood, 2023; Piper and Bagga, 2024; Hobson et al., 2024; Bamman et al., 2024) or Character Understanding and Profiling (Soni et al., 2023; Yu et al., 2023). Their capacity of memorizing important details of fictional characters has also been studied for character understanding (Stammbach et al., 2022; Zhao et al., 2024; Wang et al., 2024). In this work, we assess LLMs on the quotation attribution task systematically by accounting for memorization and annotation contamination. For this, we introduce a new measure of book memorization and show that Llama-3's state-of-the-art results are not explained by memorization but rather by its reasoning ability.

Quotation Attribution Methods to attribute direct speech to its speaker in literary texts have explored sequence labeling (O'Keefe et al., 2012), deterministic rules (Muzny et al., 2017b) or generation (Su et al., 2023). BookNLP, a popular Natural Language Processing pipeline dedicated to books, also proposes a quotation attribution system that was recently improved (Vishnubhotla et al., 2023; Michel et al., 2024). The current state-of-the-art on English novels is a recent reimplementation of BookNLP+ that uses SpanBERT (Joshi et al., 2019) as the base encoder (Michel et al., 2024).

Memorization The zero-shot and few-shot performance of LLMs has often been attributed to memorization (Lee et al., 2022; Razeghi et al., 2022a; Carlini et al., 2023). This raises important concerns in literary studies as some novels are present more often in the pretraining data of LLMs than others, creating discrepancies in downstream tasks (Chang et al., 2023). Assessing the impact of memorization on downstream tasks gives insights into LLMs capacity to generalize to unseen data, and is thus of critical importance.

Annotation Contamination Annotation contamination (Magar and Schwartz, 2022) occurs when downstream task *evaluation data* (i.e. the exact annotations) is part of the LLMs pretraining corpus. Methods such as Membership Inference Attacks (Yeom et al., 2018; Mireshghallah et al., 2022; Shi et al., 2024) have been designed to evaluate an LLM ability to generate such data instances. This causes severe issues for security and privacy (Carlini et al., 2021), but also raises questions about zero-shot performance (Li and Flanigan, 2023).

	PDNC$_1$			PDNC$_2$			Unseen		
	All	**Explicit**	**Other**	**All**	**Explicit**	**Other**	**All**	**Explicit**	**Other**
ChatGPT	71^+	-	70^+	-	-	-	-	-	-
BookNLP+	78.5 (4.0)	98.6 (1.6)	68.9 (4.4)	79.2 (10.7)	93.3 (5.7)	69.6 (10.2)	98.5	99.1	98.3
Llama-3 8b	90.6 (5.2)	94.7 (2.9)	89.1 (5.7)	88.5 (4.0)	92.8 (2.1)	85.7 (4.9)	97.9	97.5	98.4

Table 1: Quotation Attribution accuracy averaged over novels (standard deviations in parentheses) for Llama-3. We take the reported results from Su et al. (2023) for ChatGPT, and from Michel et al. (2024) for BookNLP+

3 Data

We use the Project Dialogism Novel Corpus (PDNC) (Vishnubhotla et al., 2022), which contains 28 novels published between the 19th and 20th century, resulting in 37,131 quotes annotated manually with quotation attribution. PDNC is currently the largest dataset of quotation attribution.

PDNC quotes are categorized into three types: *anaphoric* quotes, introduced with a speech verb and a pronoun or common noun, *implicit* quotes, where no narrative details about the speaker are provided and *explicit* quotes, which occur when the narrator identifies the speaker using a speech verb and a proper named-mention. Examples are given in Figure 1.

Among PDNC novels, 22 novels were released in July 2022 (PDNC$_1$), while 6 novels were added in June 2023 (PDNC$_2$). The latter subset will be crucial to test for annotation contamination since it was released after Llama-3 8b's knowledge cutoff (March 2023). Additionnaly, we fully annotated a new novel that was published after this cutoff. Following PDNC guidelines, one author annotated all quotes and a second author a subset of 5 chapters. The inter-annotator agreement, measured by Cohen's κ score, reached 97% indicating almost perfect agreement. A total of 1530 quotes were annotated. We use this recent novel to assess Llama-3's generalization ability.

4 Quotation Attribution

We divide each novel by chapters, and chunk each chapter using 4096 tokens with a stride of 1024 tokens. We modify the raw text by assigning a unique identifier to each quote starting from 1 to n, where n is the number of quotes in the chunk. We also build a character-to-alias list using the gold character-list from PDNC that we include in the prompt. Given the modified text and the list of character aliases, we prompt the model to predict the speaker of quotes $1, \dots, n$ sequentially. We use *Llama-3 8b Instruct* for all experiments, and test the 70b version on the Unseen novel as its annotations are not included in the larger model pretraining data. More details are provided in Appendix A.

Baselines We compare to Su et al. (2023) ChatGPT's (*gpt-3.5-turbo-0613*) Chain-of-Thought prompting strategy where the model is prompted with a target quote and its surrounding context. We also compare to the current state-of-the-art on PDNC (Michel et al., 2024). We use the official code to train BookNLP+ with the first cross-validation split of PDNC$_1$ that we further employ to attribute quotes in PDNC$_2$ and the unseen novel.

Evaluation We follow previous works (Vishnubhotla et al., 2023; Su et al., 2023; Michel et al., 2024), and focus on *major* and *intermediate* characters, which are characters that utter at least 10 quotes in a novel. We present attribution accuracy on *explicit* and *other* quotes, (including both *anaphoric* and *implicit* utterances) (Muzny et al., 2017b; Vishnubhotla et al., 2022). Explicit utterances occur when the narrator indicates the speaker of a quote with a speech verb and a named mention, while anaphoric quotes are introduced with a speech verb and a pronoun or common noun. When no narrative information is given about the speaker of the quote, we refer to those as implicit quotes.

Results Table 1 shows surprisingly high performance for Llama3-8b, increasing the overall attribution accuracy by up to 19 points against ChatGPT on PDNC$_1$ and 12 points against BookNLP+. This gain is due to the large performance increase when attributing non-explicit quotes, that we also see on PDNC$_2$. This suggests that Llama-3 might be able to solve complex cases of reasoning such as coreference resolution in a small context, or understanding discussion patterns.

On the Unseen novel, BookNLP+ performs slightly better than Llama-3 8b overall. When increasing the model size to 70b, the performance increases to an almost perfect accuracy, and we

	Accuracy (All)		Accuracy (Explicit)		Accuracy (Others)	
	ρ	$(\texttt{Top}_5 - \texttt{Bot}_5)$	ρ	$(\texttt{Top}_5 - \texttt{Bot}_5)$	ρ	$(\texttt{Top}_5 - \texttt{Bot}_5)$
Name-Cloze	0.15	✗	0.27*	✗	0.01	✗
CSG-Memorization	0.09	✗	0.34*	✗	0.01	✗
CSG-Reasoning	0.52*	✓	0.21	✗	0.43*	✗

Table 2: Correlations (Spearman ρ) between quotation attribution accuracy and measures of memorization (* indicates $p < 0.05$), and statistical significance at 5% from a Student t-test when testing for difference in expected attribution accuracies between top 5 most memorized books and bottom 5 least memorized books $(\texttt{Top}_5 - \texttt{Bot}_5)$.

identified only 3 wrong predictions out of 1442 quotes (note that we only consider *major* and *interemediate* characters). The larger model appears to have improved reasoning abilities, yielding better attribution. While Llama-3 shows surprising performance on both subsets of PDNC, we question if those results are due to its reasoning abilities. Thus, we analyze the impact of memorization, reasoning and annotation contamination in the next section.

5 The Impact of Memorization

The extent to which LLMs have encountered books and annotations in their training data may influence and bias their assessment on downstream tasks (Razeghi et al., 2022b; Chang et al., 2023; Li and Flanigan, 2023). We thus carry out an evaluation of book memorization and annotation contamination.

Book Memorization. We use name-cloze accuracy (Chang et al., 2023) to quantify book memorization. This methods prompts an LLM to identify a masked character name in a small passage of text. Llama-3 8b achieves a 4% average accuracy on PDNC, with 13 novels showing null accuracies. Surprisingly, we found null name-cloze accuracies for canonical works such as *The Picture of Dorian Gray* compared to reported GPT-4 accuracies of 42%. This questions name-cloze's validity for Llama-3 8b, leading us to propose a new metric: *Corrupted-Speaker-Guessing* (CSG).

We design CSG as a speaker-guessing task, providing the model with the book's title, author, a passage, and a target quote. We corrupt the passage by replacing the speaker's name with a different gender-matching name that is not used in the book. This pseudonymization approach has been used for example to build narrative-focused story embeddings (Hatzel and Biemann, 2024). When making a prediction, the LLM must decide whether to use contextual cues (*reasoning*) or rely on memorized information to identify the correct speaker, despite the misleading contextual information. More details and prompt examples are provided in Appendix B

We validate CSG in two ways. First, we follow Chang et al. (2023) and present the Spearman ρ correlation between memorization metrics and the average number of search results for 10-grams randomly sampled from a book across Google, Bing, C4, and The Pile. Significant correlations were found with all memorization measures (detailed in Appendix C). Then, we ensured that all memorization metrics returned null accuracies on the unseen novel.

Impact on Quotation Attribution We calculate Spearman ρ correlations between quotation attribution accuracy and memorization and reasoning metrics. We then identify the top 5 most and least memorized (or *reasoned* in the case of CSG-Reasoning) books and test for differences in expected quotation attribution accuracy using a Student t-test. Table 2 shows positive correlations between memorization metrics and accuracy for explicit quotes, but not over all quotes. These results suggest that book memorization does not explain Llama-3's impressive performance at attributing utterances of direct-speech, as also evidenced by high CSG-reasoning correlations. See Appendix D for detailed results per novel.

Annotation Contamination. We use `Min-K%` (Shi et al., 2024), a popular contamination detection method, with 20% randomly sampled annotation instances per novel. For each data instance, we verbalize it in plain text, and then compute `Min-K%` by averaging conditional probabilities of the `K%` tokens with the lowest values in the sequence.

A key challenge in analyzing Llama-3 probabilities is that annotation instances contain quotes and entities from novels, which can lead to variations in perplexity depending on the number of memorized passages from the book. To address this, we propose an econometrics-inspired approach:

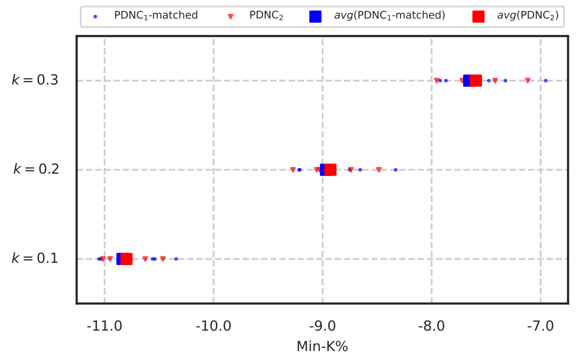

Figure 2: `Min-K%` results for various values of `K` for $PDNC_2$ and each matched novel in $PDNC_1$.

propensity score matching (Rosenbaum and Rubin, 1983) to control the influence of book memorization when analyzing Llama-3 probabilities. We begin by calculating a propensity score for each novel by fitting a logisitic regression, with the indicator of a novel being in $PDNC_2$ as the predictor. We include CSG-Memorization, name-cloze and `Min-K%` as covariates, as well as overall quotation attribution accuracy, which may vary based on whether the annotations are memorized or not. Predicted propensity scores reflect the likelihood of a novel belonging to $PDNC_2$, and hence indicate the probability that its annotations are unseen by Llama-3, given its degree of memorization. For each novel in $PDNC_2$, we match a novel in $PDNC_1$ with the closest propensity score. Figure 2 displays the average log-probabilities for each $PDNC_2$ novel and their $PDNC_1$ match. We test for differences in expected value between the `Min-K%` values with a paired Student t-test, and found no significant differences, suggesting that Llama-3 8b is unlikely to have memorized annotation instances of $PDNC_1$ (see Appendix E for a detailed analysis).

6 Conclusion

We systematically evaluate Llama-3's zero-shot performance in quotation attribution, demonstrating that a simple Chain-Of-Thought approach accurately attributes direct-speech utterances from book chapters and significantly surpasses previous state-of-the-art models by a large margin. Then, we analyze the reasons behind such performance by evaluating the impact of memorization on the downstream task. Our results suggest that neither book memorization nor annotation contamination are key factors contributing to this improvement, suggesting Llama-3 as the current best system for quotation attribution in English literature.

7 Limitations

We proposed a new, task-specific and model-specific measure of book memorization. While this measure shows a better capacity to recognize memorization than name-cloze accuracy when used with Llama-3 8b, we note that it is specific to literary texts, and that it suffers from one of the common downsides of this kind of measures: we can not be sure that instances of data have not been seen during pretraining. Some novels in our corpus exhibit non-memorization, while we know that they are part of large corpus such as The Pile or C4, indicating that we could design better tests for book memorization. Overall, we believe that the better way to test generalization of LLMs on a downstream task is to provide it with completely unseen data, which we tested by evaluating Llama-3 on a new, recently published novel.

Our metric, CSG, also labels prediction as a *reasoning* class. In reality, we can not be sure that the LLM is indeed *reasoning* as a human would do, and we instead use this specific word to indicate that the LLM is processing contextual information, and is able to prioritize this contextual information over the uncorrupted passage it has memorized. Besides, it is hard to understand why it prioritizes *reasoning* over *memorization*, and it is possible that larger models would prioritize more memorization.

The significant improvement of Llama-3 over baselines such as BookNLP+ on quotation attribution creates new possibilities to better analyze large corpora of literary texts. However, this improvement comes with longer inference times, taking up to a GPU hour for a single novel and limiting its impact for the study of massive corpora such as Project Gutenberg. In comparison, BookNLP+ makes predictions in a few minutes for a novel.

In this work, we prompted Llama-3 with a predefined gold character-to-alias list. In real-world scenarios, this list is unlikely to be available. Although approaches to build an alias list have been widely explored in the literature, our work does not mirror the full workflow of character discovery followed by quotation attribution.

References

David Bamman, Kent K. Chang, Li Lucy, and Naitian Zhou. 2024. On classification with large language models in cultural analytics. *Preprint*, arXiv:2410.12029.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. *Preprint*, arXiv:2202.07646.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. *Preprint*, arXiv:2012.07805.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Hans Ole Hatzel and Chris Biemann. 2024. Story embeddings — narrative-focused representations of fictional stories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5931–5943, Miami, Florida, USA. Association for Computational Linguistics.

David G Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032, Miami, Florida, USA. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys*, 52(5):1–40.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *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.

Changmao Li and Jeffrey Flanigan. 2023. Task contamination: Language models may not be few-shot anymore. *Preprint*, arXiv:2312.16337.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Gaspard Michel, Elena V. Epure, Romain Hennequin, and Christophe Cerisara. 2024. Improving quotation attribution with fictional character embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12723–12735, Miami, Florida, USA. Association for Computational Linguistics.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky. 2017a. Dialogism in the novel: A computational model of the dialogic nature of narration and quotations. *Digital Scholarship in the Humanities*, 32:ii31–ii52.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017b. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain. Association for Computational Linguistics.

Timothy O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799, Jeju Island, Korea. Association for Computational Linguistics.

Andrew Piper and Sunyam Bagga. 2024. Using large language models for understanding narrative discourse. In *Proceedings of the The 6th Workshop on Narrative Understanding*, pages 37–46, Miami, Florida, USA. Association for Computational Linguistics.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022a. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022b. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. *Preprint*, arXiv:2310.16789.

Matthew Sims and David Bamman. 2020. Measuring information propagation in literary social networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 642–652, Online. Association for Computational Linguistics.

Sandeep Soni, Amanpreet Sihra, Elizabeth Evans, Matthew Wilkens, and David Bamman. 2023. Grounding characters and places in narrative text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11723–11736, Toronto, Canada. Association for Computational Linguistics.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Zhenlin Su, Liyan Xu, Jin Xu, Jiangnan Li, and Mingdu Huangfu. 2023. Sig: Speaker identification in literature via prompt-based generation. *Preprint*, arXiv:2312.14590.

Ted Underwood. 2023. Using gpt-4 to measure the passage of time in fiction.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The project dialogism novel corpus: A dataset for quotation attribution in literary texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848, Marseille, France. European Language Resources Association.

Krishnapriya Vishnubhotla, Frank Rudzicz, Graeme Hirst, and Adam Hammond. 2023. Improving automatic quotation attribution in literary novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 737–746, Toronto, Canada. Association for Computational Linguistics.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Qian Wang, and Yue Zhang. 2024. Novelqa: A benchmark for long-range novel question answering. *Preprint*, arXiv:2403.12766.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. *Preprint*, arXiv:1709.01604.

Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023. Personality understanding of fictional characters during book reading. *Preprint*, arXiv:2305.10156.

Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, and Lin Gui. 2024. Narrative-Play: Interactive narrative understanding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 82–93, St. Julians, Malta. Association for Computational Linguistics.

A Method Details - Quotation Attribution

We divide novels in chapters, and build chunks of text of length 4096 tokens with a stride of 1024 tokens. If an entire chapter is less than 4096 tokens, then we use all tokens in this chapter and do not use striding for the next chunk. That is we only use striding when chapters are longer than 4096 tokens. All quotes in a chunk need to be predicted by the model.

With the above chunk construction, some quotes will be predicted twice when striding is used. We experiment with two approaches:

1. We consider only the first prediction of a quote, i.e. the first time it appears in a chunk.
2. We propose an incremental prompting strategy, where predictions of overlapping quotes are also given as contextual information, and we prompt the LLM to predict all quotes in a chunk, refining its prediction if necessary.

In all cases, we use Chain-of-Thought prompting, and prompt the model with the gold character-to-alias list. We tested without using this list, but we realized that the model was often predicting aliases that were not in this list, which made the attribution to a character ID a lot harder. We found that using the gold character-to-alias list is the most straightforward way to restrict the generation to a candidate name, but also makes our results an upper-bound when evaluating the end-to-end workflow of quotation attribution that also includes building a silver character-to-alias list. Note that the gold character list is also used by other baselines (ChatGPT and BookNLP+), making the comparison with our approach still fair.

A prompt example used in strategy (1.) is displayed in Figure 7 and an incremental prompt example used when there are overlapping quotes in strategy (2.) is displayed in Figure 8.

The model output is a JSON string, with unique quote identifiers as keys and predicted names as values. In particular, we use the character-to-alias list to replace the predicted name with their canonical character ID (which is our gold label). If the model generates a name that is not an alias, we consider its predictions as wrong (*i.e.* we do not use any lenient metrics such as substring matching).

Results for both strategies on $PDNC_2$ are displayed in Table 3. We found that the incremental strategy led to slightly better results on this subset of novels, and thus used it for all experiments.

	All	**Explicit**	**Others**
Strategy 1.	87.6 (3.9)	92.0 (2.5)	84.7 (4.9)
Strategy 2.	88.5 (4.0)	92.8 (2.1)	85.7 (4.9)

Table 3: Average Quotation Attribution accuracy on $PDNC_2$, with (standard deviation) for both strategies.

B Method Details - CSG

We designed *Corrupted-Speaker-Guessing* by finding out the really low/null name-cloze accuracies of Llama-3 8b on PDNC. These results suggests that Llama-3 has not *exactly memorized* some canonical PDNC novels. To avoid a similar situation where CSG returns null accuracies, we also provide book-level metadata as contextual information to be able to catch *weaker memorization*. CSG prompts an LLM with a corrupted passage of a book, the book's title and author, and a target quote appearing in the passage. The passage contains 10 sentences before and after the target quote (we use SpaCy to segment sentences). It tasks the LLM to find the speaker of the target quote. To corrupt the original passage, we apply the following modifications:

1. We find all proper named mentions of the speaker, using the gold character-to-alias list.
2. We replace all proper named mentions of the speaker with another name, matching its gender. We use two first names for each gender: "Henry" or "Joseph" and "Emma" or "Elizabeth". We also use three last names: "Stone", "Walker" and "Smith". We use combinations of first and last names such that none of these names appear in the novel. Finally, we kept all honorifics when replacing ("Miss Bates" $\rightarrow$ "Miss Smith").

Note that this process was done manually by one of the author and that we never used "Emma Stone" or other celebrity names that are likely to appear more frequently on the web.

We use two different prompts, depending on whether the target quote is an explicit quote or non-explicit. In the case of explicit quotes, we formulate the task as a cloze, replacing all named mentions and masking the referring expression ("said [MASK]"). An example is provided in Figure 3. For other quote types, we do not use masking and use the prompt provided in Figure 5 and Figure 6.

	Google	Bing	C4	Pile
Name-Cloze	0.42	0.55	0.75	0.57
CSG-Mem	0.54	0.3	0.42	0.61
`Cloze Only`	0.65	0.44	0.45	0.53

Table 4: Correlation (Spearman ρ) between Llama-3 memorization measures and number of search results in Google, Bing, C4 and the Pile. All coefficients are significative except for CSG-Mem and Bing.

We ensure that there is at least one named mention of the speaker in the corrupted passage, such that contextual information should point to the corrupted character name as the speaker.

For each quote type (explicit, anaphoric and implicit), we randomly sample 100 quotes and their associated corrupted passages, and prompt the model to find the speaker of the target quote. Given the model's prediction, we calculate two types of accuracy:

- Memorization accuracy: when the model predicts the true speaker name, even though the passage does not contain any named mention of this speaker.
- Reasoning accuracy: when the model uses contextual information to predict the corrupted speaker name.

We calculate CSG-Memorization and CSG-Reasoning accuracies by averaging each accuracy over all quote types.

C CSG Validation

One of the validation of CSG was done following (Chang et al., 2023), by evaluating the correlation between (a proxy of) the frequency of of a novel on the web and its memorization accuracy. We present in Table 4 all correlation results between the average number of search results of random 10-grams on different databases, and memorization metrics. We do not have access to the custom search APIs that were used in Chang et al. (2023), so we instead directly use their reported number of searches for each endpoint. We gathered data for a subset of 16 PDNC novels that were also used by (Chang et al., 2023), and calculate Spearman ρ correlations between the memorization measures and the average number of search results.

D Results per Novel for CSG and Name-Cloze

We present in Table 5 all memorization and reasoning accuracies. We also chose to display the CSG-Memorization accuracy with the cloze prompt (with explicit quotes) as it holds interesting properties: we found similar conclusions when replacing CSG-Memorization with the cloze variant of CSG-Memorization. This cloze variant is more practical, as automatically finding speakers of explicit quotes in novels is usually the easiest attribution task among all quote types, as shown by all systems accuracy. Therefore, one can use only CSG-Memorization Cloze as a measure of book memorization, removing the need for annotating all quote types to measure the full CSG-Memorization.

E Annotation Contamination Results per Novel

We calculate `Min-K%` by verbalizing instances of data. We present in Figure 4 an example of how we verbalize an instance of data. We then calculate the conditional log-probabilities of each token in the verbalized sequence, and average the $k\%$ lowest log-probabilities in the sequence, for $k = 10, 20, 30$.

Given each novel in PDNC_2 and their PDNC_1 match, we conduct a paired paired Student t-test and test for difference in expected `Min-K%` values. We found no statistical differences ($t = 0.54$, $p = 0.3$).

Other approaches to detect contamination involves a chronological analysis (Li and Flanigan, 2023), comparing downstream performance on a set of data that is known to be inside the pretraining corpus to the performance on a set not included during pretraining. We follow the same approach as described in the Annotation Contamination paragraph of Section 5, but instead define the outcome variable to be the quotation attribution accuracy rather than Min-K% when matching with propensity score. We found no significant differences in the expected values of quotation attribution accuracy ($t = 0.75$, $p = 0.25$) using a paired t-test from matched novels.

F Computing Information

We used a 32-core Intel Xeon Gold 6244 CPU @ 3.60GHz CPU with 128GB RAM equipped with 3 RTX A5000 GPUs with 24GB RAM. We used a single RTX A5000 for all `Llama3-8b` experiments. We used the 8-bits version of `Llama3-8b-Instruct` using the *BitsAndBytes* library. The peak memory used was around 14GB of RAM. We employ a relatively large contextual window, and ask the model to generates long attribution lists. Thus, we observed quite large inference times, and processing entire novels varied from 10 minutes to an hour. For the `Llama3-70b` experiments, we used one A100-80GB and used the 4-bits quantized version `Meta-Llama-3-70B-Instruct-Q4_K_M.gguf`.

> You will be given a passage of the book Persuasion written by Jane Austen that you have seen in your training data. Find the proper name that fills the [MASK] token. This name is a proper name (not a pronoun or any other word). You must make a guess, even if you are uncertain. Do not explain your reasoning.
>
> You must format your answer in <speaker>[SPEAKER]< speaker> tags.
>
> Passage:
>
> [. . .]
>
> "It was my friend Mrs Rooke; Nurse Rooke; who, by-the-bye, had a great curiosity to see you, and was delighted to be in the way to let you in. She came away from Marlborough Buildings only on Sunday; and she it was who told me you were to marry Mr Elliot. She had had it from Mrs Wallis herself, which did not seem bad authority. She sat an hour with me on Monday evening, and gave me the whole history." **"The whole history," repeated [MASK]**, laughing. "She could not make a very long history, I think, of one such little article of unfounded news."
>
> Mrs Smith said nothing.
>
> "But," **continued Emma**, presently, "though there is no truth in my having this claim on Mr Elliot, I should be extremely happy to be of use to you in any way that I could. Shall I mention to him your being in Bath? Shall I take any message?"
>
> [. . .]
>
> Target quote:
> "The whole history,"

Figure 3: Example of a CSG prompt with an explicit quote. Here, the character *Anne Elliot* from *Persuasion* is replaced by *Emma*.

> **Raw Data:** "Q0","and what is the use of a book, without pictures or conversations?","[ánd what is the use of a book,, ẃithout pictures or conversations?]́","[[254, 284], [301, 335]]","Alice","[]","Explicit","thought Alice","[[], []]","[[], []]","[[], []]"
>
> **Verbalized Data:** quoteID: Q0; quoteText: and what is the use of a book, without pictures or conversations?; subQuotationList: ['and what is the use of a book,', 'without pictures or conversations?']; quoteByteSpans: [[254, 284], [301, 335]]; speaker: Alice; addressees: []; quoteType: Explicit; referringExpression: thought Alice; mentionTextsList: [[], []]; mentionSpansList: [[], []]; mentionEntitiesList: [[], []]

Figure 4: Example of a verbalized instance of data.

> You will be given a passage of the book Persuasion written by Jane Austen that you have seen in your training data. Find the true speaker name of the target quote. This name is a proper name (not a pronoun or any other word). You must make a guess, even if you are uncertain. Do not explain your reasoning.
>
> You must format your answer in <speaker>[SPEAKER]< speaker> tags.
>
> Passage:
>
> [. . .]
>
> **Captain Stone** left his seat, and walked to the fire-place; probably for the sake of walking away from it soon afterwards, and taking a station, with less bare-faced design, by Anne.
>
> "You have not been long enough in Bath," said he, "to enjoy the evening parties of the place."
>
> "Oh! no. The usual character of them has nothing for me. I am no card-player."
>
> "You were not formerly, I know. You did not use to like cards; but time makes many changes."
>
> "I am not yet so much changed," cried Anne, and stopped, fearing she hardly knew what misconstruction. After waiting a few moments he said, and as if it were the result of immediate feeling, "It is a period, indeed! Eight years and a half is a period."
>
> [. . .]
>
> Target quote:
>
> "You were not formerly, I know. You did not use to like cards; but time makes many changes."

Figure 5: Example of a CSG prompt with an implicit quote. Here, the character *Captain Wentworth* from *Persuasion* is replaced by *Captain Stone*.

Title	Author	Name-Cloze	CSG-M	CSG-M (Cloze)	CSG-R
The Age of Innocence	Edith Wharton	0.0	0.27	0.27	0.5
Pride and Prejudice	Jane Austen	0.1	0.23	0.27	0.59
The Picture Of Dorian Gray	Oscar Wilde	0.0	0.22	0.44	0.48
The Awakening	Kate Chopin	0.0	0.21	0.28	0.49
Emma	Jane Austen	0.19	0.2	0.24	0.55
Daisy Miller	Henry James	0.0	0.19	0.46	0.7
A Room With A View	E. M. Forster	0.0	0.17	0.24	0.53
The Sun Also Rises	Ernest Hemingway	0.01	0.17	0.34	0.5
Sense and Sensibility	Jane Austen	0.04	0.16	0.16	0.7
Northanger Abbey	Jane Austen	0.03	0.12	0.2	0.64
Anne Of Green Gables	Lucy M. Montgomery	0.02	0.12	0.3	0.75
Alice's Adventures in Wonderland	Lewis Carroll	0.47	0.12	0.27	0.61
Persuasion	Jane Austen	0.0	0.11	0.21	0.62
The Sign of the Four	Arthur Conan Doyle	0.03	0.06	0.08	0.34
The Invisible Man	Herbert George Wells	0.02	0.06	0.16	0.88
Howards End	Edward Morgan Forster	0.0	0.05	0.09	0.53
The Mysterious Affair At Styles	Agatha Christie	0.0	0.03	0.06	0.63
A Handful Of Dust	Evelyn Waugh	0.0	0.02	0.0	0.57
The Gambler	F. M. Dostoevsky	0.01	0.02	0.04	0.58
Night and Day	Virginia Woolf	0.0	0.01	0.03	0.78
The Man Who Was Thursday	Gilbert K. Chesterton	0.0	0.0	0.0	0.67
The Sport of the Gods	Paul Laurence Dunbar	0.0	0.0	0.0	0.64
A Passage to India	Edward Morgan Forster	0.0	0.12	0.17	0.43
Mansfield Park	Jane Austen	0.0	0.09	0.13	0.59
Winnie-The-Pooh	Alan Alexander Milne	0.06	0.07	0.14	0.66
Where Angels Fear to Tread	Edward Morgan Forster	0.0	0.04	0.08	0.57
Oliver Twist	Charles Dickens	0.07	0.03	0.06	0.71
Hard Times	Charles Dickens	0.02	0.01	0.01	0.78
Dark Corners	Katie Rush	0.0	0.0	0.0	0.84

Table 5: All Memorization and Reasoning accuracies calculated with Llama-3 8b per novel. Top: $PDNC_1$, Middle: $PDNC_2$, Bottom: Unsenn novel.

You will be given a passage of the book Persuasion written by Jane Austen that you have seen in your training data. Find the true speaker name of the target quote. This name is a proper name (not a pronoun or any other word). You must make a guess, even if you are uncertain. Do not explain your reasoning.

You must format your answer in <speaker>[SPEAKER]< speaker> tags.

Passage:

[. . .]

Charles shewed himself at the window, all was ready, their visitor had bowed and was gone, the Miss Musgroves were gone too, suddenly resolving to walk to the end of the village with the sportsmen: the room was cleared, and **Emma might finish her breakfast as she could**.

"It is over! it is over!" **she repeated to herself** again and again, in nervous gratitude. "The worst is over!"

[. . .]

Target quote:

"The worst is over!"

Figure 6: Example of a CSG prompt with an anaphoric quote. Here, the character *Anne Elliot* from *Persuasion* is replaced by *Emma*.

Instruction: You are an excellent linguist working in the field of literature. I will provide you with a passage of a book where some quotes have unique identifiers marked by headers '|quote_id|'. Your are tasked to build a list of quote attributions by sequentially attributing the marked quotes to their speaker.

Passage:

—

Chapter 8

From this time Captain Wentworth and Anne Elliot were repeatedly in the same circle. They were soon dining in company together at Mr Musgrove's, for the little boy's state could no longer supply his aunt with a pretence for absenting herself; and this was but the beginning of other dinings and other meetings.

Whether former feelings were to be renewed must be brought to the proof; former times must undoubtedly be brought to the recollection of each; they could not but be reverted to; the year of their engagement could not but be named by him, in the little narratives or descriptions which conversation called forth. His profession qualified him, his disposition lead him, to talk; and |1|"That was in the year six;"|1| |2|"That happened before I went to sea in the year six,"|2| occurred in the course of the first evening they spent together: and though his voice did not falter, and though she had no reason to suppose his eye wandering towards her while he spoke, Anne felt the utter impossibility, from her knowledge of his mind, that he could be unvisited by remembrance any more than herself. There must be the same immediate association of thought, though she was very far from conceiving it to be of equal pain.

[. . .]

|50|"Aye, to be sure. Yes, indeed, oh yes! I am quite of your opinion, Mrs Croft,"|50| was Mrs Musgrove's hearty answer. |51|"There is nothing so bad as a separation. I am quite of your opinion. I know what it is, for Mr Musgrove always attends the assizes, and I am so glad when they are over, and he is safe back again."|51|

The evening ended with dancing. On its being proposed, Anne offered her services, as

—

Step 1: Attribute sequentially each quote to their speaker.
Step 2: Match each speaker found in the previous step with one of the following name:
Names

—

Admiral Croft=The Admiral=Admiral
Anne Elliot=Miss Anne=Miss Anne Elliot=Anne
Captain Harville=Harville
Captain Wentworth=Wentworth=Frederick Wentworth=Frederick
Charles Hayter=Hayter
Charles Musgrove
Elizabeth
Henrietta Musgrove=Henrietta
Lady Dalrymple=Dalrymple
Lady Russell=Russell
Louisa Musgrove=Louisa
Mary Musgrove=Mary
Mr Shepherd=Shepherd=John Shepherd
Mrs Clay=Clay=Penelope
Mrs Musgrove=Musgrove
Mrs Smith=Hamilton=Smith=Miss Hamilton
Sir Walter Elliot=Walter Elliot=Sir Walter=Walter
Sophia Croft=Sister Of Captian Wentworth=Croft=Mrs Croft
The Waiter=Waiter
William Walter Elliot=William=Mr Elliot=Elliot

—

Step 3: Replace the speakers found in Step 1 with their matching name found in Step 2. Your answer should follow this JSON format:
{
'quote_id_1' : 'predicted_speaker_1',
'quote_id_2' : 'predicted_speaker_2'
}
Your answer should only contain the output of **Step 3** and can only contain quote identifiers and speakers. Never generate quote content and don't explain your reasoning.

Figure 7: Example of a prompt used when there are no overlapping quotes. We also only use this prompt when experiment without incremental updating. The novel here is *Persuasion*.

Instruction: You are an excellent linguist working in the field of literature. I will provide you with a passage of a book where some quotes have unique identifiers marked by headers '|quote_id|'. You will also be provided a list of characters and their aliases, and previous predictions. Your are tasked to build a list of quote attributions by sequentially attributing the marked quotes to their speaker.

Passage:

—

|1|"then?"|1|

|2|"All merged in my friendship, Sophia. I would assist any brother officer's wife that I could, and I would bring anything of Harville's from the world's end, if he wanted it. But do not imagine that I did not feel it an evil in itself."|2|

|3|"Depend upon it, they were all perfectly comfortable."|3|

|4|"I might not like them the better for that perhaps. Such a number of women and children have no right to be comfortable on board."|4|

[. . .]

|19|"I beg your pardon, madam, this is your seat;"|19| and though she immediately drew back with a decided negative, he was not to be induced to sit down again.

Anne did not wish for more of such looks and speeches. His cold politeness, his ceremonious grace, were worse than anything.

—

Previous predictions:

—

{ '2': 'pred_0', '4': 'pred_1', '6': 'pred_2', '11': 'pred_3', '12': 'pred_4' }

—

Step 1: Attribute sequentially each quote to their speaker. Update the previous predictions if you think it contains wrong speaker prediction.
Step 2: Match each speaker found in the previous step with one of the following name:
Names

—

Admiral Croft=The Admiral=Admiral
Anne Elliot=Miss Anne=Miss Anne Elliot=Anne
Captain Harville=Harville
Captain Wentworth=Wentworth=Frederick Wentworth=Frederick
Charles Hayter=Hayter
Charles Musgrove
Elizabeth
Henrietta Musgrove=Henrietta
Lady Dalrymple=Dalrymple
Lady Russell=Russell
Louisa Musgrove=Louisa
Mary Musgrove=Mary
Mr Shepherd=Shepherd=John Shepherd
Mrs Clay=Clay=Penelope
Mrs Musgrove=Musgrove
Mrs Smith=Hamilton=Smith=Miss Hamilton
Sir Walter Elliot=Walter Elliot=Sir Walter=Walter
Sophia Croft=Sister Of Captian Wentworth=Croft=Mrs Croft
The Waiter=Waiter
William Walter Elliot=William=Mr Elliot=Elliot

—

Step 3: Replace the speakers found in Step 1 with their matching name found in Step 2. Your answer should follow this JSON format:
{
'quote_id_1' : 'predicted_speaker_1',
'quote_id_2' : 'predicted_speaker_2'
}
Your answer should only contain the output of **Step 3** and can only contain quote identifiers and speakers. Never generate quote content and don't explain your reasoning.

Figure 8: Example of an incremental prompt used when there are overlapping quotes between the last chunk and the current chunk. The novel here is *Persuasion*.