On the Limitations of Large Language Models (LLMs): False Attribution

Tosin Adewumi[*], Nudrat Habib, Lama Alkhaled & Elisa Barney
ML Group, EISLAB, Luleå University of Technology, Sweden.
`{firstname.lastname}@ltu.se`

Abstract

In this work, we provide insight into one important limitation of parametric knowledge of large language models (LLMs), i.e. false attribution, and introduce a new hallucination metric - Simple Hallucination Index (SHI). We empirically evaluate the power of 3 open SotA LLMs in zero-shot setting (LLaMA-2-13B, Mixtral 8x7B, and Gemma-7B), especially as human annotation can be costly. The average results show that Mixtral 8x7B has the highest prediction accuracy, the lowest SHI, and a Pearson's correlation (r) of 0.737, 0.249, and -0.9996, respectively, followed by LLaMA-2-13B and Gemma-7B. However, Mixtral 8x7B suffers from high hallucinations for 3 books, rising as high as an SHI of 0.87 (in the range 0-1, where 1 is the worst). The other two models also do. We publicly release the annotated chunks of data and our codes for further exploration.

1 Introduction

Prompt Input:	Who is the author of this text: *'how–does the young gentleman stand for Haverford West?...he wheeled about; and going round the levee, spoke to every individual, with the most courteous familiarity;...As for the common business of the nation, it is carried on in a constant routine by the clerks of the different offices, otherwise the wheels of government'*
Mixtral 8x7B Output:	The text you provided is an excerpt from a longer work, and it does not include any information about the author. The text appears to be a piece of fiction, describing a scene in which a duke is holding a levee and receiving visitors. **The author of this text is unknown in this context**.

Table 1: Output example of Mixtral 8x7B with input from a fragment of *The Expedition of Humphry Clinker* by Tobias Smollett. (Bold style for emphasis.)

False attribution is the incorrect representation that someone or an entity is the author of a work when they are actually not (Carty and Hodkinson, 1989). this problem raises ethical , moral and legal issues. Hallucination, in the context of AI, is when a model confidently presents false information as fact (Maynez et al., 2020; Ji et al., 2023). Due to the high cost of human annotation, it is appealing to use automatic annotation by LLMs, which are large neural probabilistic models that are pretrained on large amounts of data (**including books**) through self-supervised learning to predict the next token and finetuned for downstream tasks (Radford et al., 2019; Brown et al., 2020; Adewumi et al., 2023). It appears many existing hallucination metrics are based on a binary format, such as factual or non-factual (Lee et al., 2022; Kang et al., 2024), yes or no,[1] and other binary options (Li et al., 2023). This is inadequate and misleading, especially for a task such as Question Answering (QA), as we believe a system should not be penalized for saying *I don't know*, as in the example in Table 1

In this work, our objective is to demonstrate, in **zero-shot setting**, the strengths and limitations of LLMs with regards to the task of author attribution for chunks of text and introduce a simple hallucination metric for their evaluation - Simple Hallucination Index (SHI). In order to answer our research question of "**how do recent open LLMs fare with regards to false attribution for short texts of books?**", we selected the 10 most downloaded (or popular) books which are provided in Table 2,[2] according to Project Gutenberg. More details about the books are provided in Section 3.

Our contributions include the following:

- We introduce a simple and novel hallucination metric for LLMs - Simple Hallucination Index (SHI) (pronounced *shy*). This is important to build more trustworthy GenAI.

[1] docs.rungalileo.io/galileo/gen-ai-studio-products/guardrail-store/factuality

[2] for the month of March, 2024; at gutenberg.org/ebooks/bookshelf

- We publicly release the LLM-annotated chunks of data, which can be useful for author attribution tasks[3].
- We are the first, to the best of our knowledge, to demonstrate the false attribution problem in LLMs in a systematic way for chunks of books?

The rest of this paper is organized as follows. In Section 2, we explain the SHI metric. Section 3 discusses the methods. We present the results and analysis in Section 4 and conclusion in Section 5.

2 Simple Hallucination Index (SHI)

SHI, given by Equation 1, differentiates unknown (u) from incorrect (i) facts made by an LLM, unlike the typical binary (correct/incorrect) classes in author attribution tasks (Diederich et al., 2003; Savoy, 2016) or hallucination metrics. A binary metric takes the form of Equation 2 and is too restrictive. It forces an exaggeration of the evaluation, where the incorrect (i^*) is a combination of the actual incorrect and the unknown cases. The correct predictions are represented by c in both equations.

$$SHI = \frac{i}{c+i+u} \tag{1}$$

$$Binary = \frac{i^*}{c+i^*} = \frac{i+u}{c+i+u} \tag{2}$$

This important property of SHI, in considering the unknown (when the model is unable to give any prediction or explicitly says it's unsure), ensures it does not score the model positively. This contrasts with the *truthfulness* metric of TruthfulQA (Lin et al., 2022) that assigns a score even when the model refuses to answer a question for any reason, the *ensemble* of FactualityPrompt (Lee et al., 2022) that is binary-based on factual and non-factual annotations, and HaluEval's accuracy (Li et al., 2023), which is also binary-based on hallucinated or normal samples. Furthermore, these metrics are tied to specific benchmarks or data about world facts, making them less flexible. On the other hand, SHI can be applied to any task involving LLMs and is not dependent on any specific benchmark or dataset.

If we compare SHI to other metrics like Precision, recall, F1, accuracy and *Metric for Evaluation of Translation with Explicit ORdering (METEOR)* which may be used in hallucination evaluation (Chen et al., 2023; Chang et al., 2024), we can observe their limitation. This is because such metrics are based on true positives (*tp*), true negatives (*tn*), false positives (*fp*), and false negatives (*fn*), none of which accounts for unknown cases.

[3]available after anonymity period

3 Methodology

All the experiments were performed on an Nvidia DGX-1 node, with 8 x 40GB A100 GPUs, that runs Ubuntu 22.04. The 3 LLMs we evaluated are chat (or instruction-tuned) models of the Large Language Model Meta AI (LLaMA)-2-13B, Mixtral 8x7B, and Gemma-7B-In. We kept the default hyper-parameters and set the maximum number of tokens for each to 1,200. We follow previous work and use accuracy to report prediction performance (Luyckx and Daelemans, 2008; Mallen et al., 2023). The 10 most downloaded (or popular) books (according to Project Gutenberg) used in this study are provided in Table 2[4]. We follow Bevendorff et al. (2019) and Hicke and Mimno (2023) and split each book into chunks of text of 400 words. The last chunk for each book usually contains less than 400 words.

3.1 Annotation by LLMs

Similar to the annotation guideline for several case studies by Ide (2017), our annotation lifecycle starts with creating the chunks from the books. We then prompt the LLMs for author attribution in a 3-fold loop, depending on if the output is empty, which occurred only with LLaMA-2. After each iteration, the prompt is redesigned before it is fed to the LLM according to the following points, where *txt* is the chunk of text. The 2 follow-up prompts are designed with instruction because of the potential to improve performance, as shown in the literature (Wei et al., 2022; Kojima et al., 2022; Adewumi et al., 2024).

1. Who is the author of this text: *'txt'*?
2. ### Instruction: Following is a Question Answering task. As a helpful system, give a suitable response: Who is the author of this text: *'txt'*?
3. ### Instruction: Following is a Question Answering task. As a helpful system, give a suitable response: Who wrote this text: *'txt'*?

[4]where P. Year: Publication Year

Table 2: The 10 most popular books according to Project Gutenberg

Book	Author	Chunks	Downloads
Pride and Prejudice	Jane Austen	306	77,172 3
Moby Dick	Herman Melville	530	69,342
Middlemarch	George Eliot	790	50,920
The Adventures of Ferdinand Count Fathom	T. Smollett	397	39,848
The Expedition of Humphry Clinker	T. Smollett	371	38,788
The Adventures of Roderick Random	T. Smollett	477	38,561
History of Tom Jones	Henry Fielding	864	37,986
A Doll's House	Henrik Ibsen	67	29,637
Crime and Punishment	Fyodor Dostoevsky	507	23,269
Great Expectations	Charles Dickens	922	19,251
Totals		5,231	424,774

After annotation, 162 chunks are randomly selected from each LLM-annotated set of chunks for human evaluation and post-processing, based on the error margin of 7% and a confidence interval of 95% for the book with the most chunks (Great Expectations). The post-processing refers to condensing the descriptive output into one word: 1) the last name of the correct author, 2) '*others*', when it's an incorrect attribution, or 3) '*unknown*', when the LLM does not know or there's still no output after the 3 prompts. Effectively, these are the 3 labels. Only LLaMA-2-13B used the additional 2nd and 3rd prompts because of the occasional empty outputs in a previous loop iteration.

4 Results and Discussion

Table 3 provides detailed results with Mixtral 8x7B having the best average performance across all scores, resulting in the best average accuracy and the lowest average SHI. Gemma-7B has the lowest average accuracy and the highest average SHI. The performance of the LLMs seem to follow the trend of their parameter sizes. The Pearson's correlation (r) values are statistically significant, based on p-value < 0.00001 for alpha of 0.05 for all the models. We observe, based on SHI, that it is better for a model to admit it does not know an answer than to make a false attribution. We also observe a strong negative correlation between accuracy and SHI, based on r, which is indicative of the fidelity of SHI in effectively scoring hallucinations. Despite having the best average performance, Mixtral 8x7B hallucinates strongly on all the 3 books by *Smollett*. This issue is observed for all the LLMs. Figure 1 depicts the 3 metrics for Mixtral 8x7B.

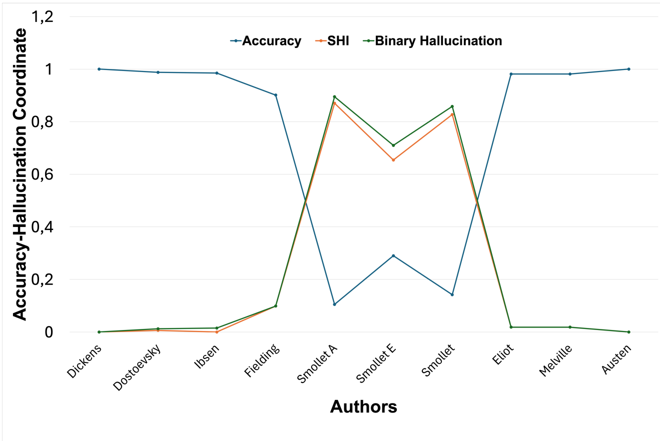

Figure 1: Correlation of accuracy, SHI, and binary hallucination for Mixtral 8x7B..

5 Conclusion

We showed, in this work, that recent LLMs are powerful but they still suffer from high hallucinations in some cases when it comes to author attribution. Our newly introduced SHI, the hallucination metric, demonstrates fidelity in providing an effective score for hallucination in a given task. This new metric has a strong negative correlation with prediction accuracy. We strongly believe that adequately gauging a problem will provide the opportunity to more adequately tackle it.

Table 3: Detailed results

Ground Truth	Model	Acc ↑	# Correct	# Others	# Unknown	SHI ↓	Binary Hallucination ↓
Austen	LLaMA-2-13	0.586	95	3	64	0.019	0.413
	Mixtral 8x7B	1	162	0	0	0	0
	Gemma-7B	0.765	124	3	35	0.019	0.234
Melville	LLaMA-2-13	0.667	108	2	52	0.012	0.333
	Mixtral 8x7B	0.981	159	3	0	0.019	0.018
	Gemma-7B	0.580	94	21	47	0.130	0.419
Eliot	LLaMA-2-13	0.611	99	24	39	0.148	0.388
	Mixtral 8x7B	0.981	159	3	0	0.019	0.018
	Gemma-7B	0.086	14	72	76	0.444	0.913
Smollett	LLaMA-2-13	0.025	4	113	45	0.698	0.965
	Mixtral 8x7B	0.142	23	134	5	0.827	0.858
	Gemma-7B	0	0	41	121	0.253	1
Smollett (Expedition)	LLaMA-2-13	0.012	2	116	44	0.716	0.987
	Mixtral 8x7B	0.290	47	106	9	0.654	0.709
	Gemma-7B	0	0	88	74	0.543	1
Smollett (Adventures of Roderick)	LLaMA-2-13	0.006	1	116	45	0.716	0.993
	Mixtral 8x7B	0.105	17	141	4	0.870	0.895
	Gemma-7B	0	0	88	74	0.543	1
Fielding	LLaMA-2-13	0.395	64	44	54	0.272	0.604
	Mixtral 8x7B	0.901	146	16	0	0.098	0.098
	Gemma-7B	0.025	4	80	78	0.494	0.975
[5]Ibsen	LLaMA-2-13	0.493	33	2	32	0.030	0.507
	Mixtral 8x7B	0.985	66	0	1	0	0.014
	Gemma-7B	0.552	37	29	1	0.433	0.447
Dostoevsky	LLaMA-2-13	0.617	100	6	56	0.037	0.382
	Mixtral 8x7B	0.988	160	1	1	0.006	0.012
	Gemma-7B	0.741	120	14	28	0.086	0.259
Dickens	LLaMA-2-13	0.815	132	1	29	0.006	0.185
	Mixtral 8x7B	1	162	0	0	0	0
	Gemma-7B	0.463	75	47	40	0.290	0.537

References

Tosin Adewumi, Lama Alkhaled, Claudia Buck, Sergio Hernandez, Saga Brilioth, Mkpe Kekung, Yelvin Ragimov, and Elisa Barney. 2023. Procot: Stimulating critical thinking and writing of students through engagement with large language models (llms). *arXiv preprint arXiv:2312.09801*.

Tosin Adewumi, Nudrat Habib, Lama Alkhaled, and Elisa Barney. 2024. Instruction makes a difference. *arXiv preprint arXiv:2402.00453*.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2019. Generalizing unmasking for short texts. 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)*, pages 654–659.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hazel Carty and Keith Hodkinson. 1989. Copyright, designs and patents act 1988. *The Modern Law Review*, 52(3):369–379.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 245–255.

Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2003. Authorship attribution with support vector machines. *Applied intelligence*, 19:109–123.

Rebecca M. M. Hicke and David Mimno. 2023. T5 meets tybalt: Author attribution in early modern english drama using large language models. In *Computational Humanities Research Conference 2023, Proceedings*, pages 274–302.

Nancy Ide. 2017. *Introduction: The handbook of linguistic annotation*. Springer.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Comparing hallucination detection metrics for multilingual generation. *arXiv preprint arXiv:2402.10496*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd international conference on computational linguistics (COLING 2008)*, pages 513–520.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jacques Savoy. 2016. Estimating the probability of an authorship attribution. *Journal of the Association for Information Science and Technology*, 67(6):1462–1472.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.