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

Tosin Adewumi*, Nudrat Habib, Lama Alkhaled, and Elisa Barney
Machine Learning Group, EISLAB,
Luleå University of Technology, Sweden
firstname.lastname@ltu.se

Abstract

In this work, we introduce a new hallucination metric - Simple Hallucination Index (SHI) and provide insight into one important limitation of the parametric knowledge of large language models (LLMs), i.e. false attribution. The task of automatic author attribution for relatively small chunks of text is an important NLP task but can be challenging. We empirically evaluate the power of 3 open SotA LLMs in zero-shot setting (Gemma-7B, Mixtral 8x7B, and LLaMA-2-13B). We acquired the top 10 most popular books of a month, according to Project Gutenberg, divided each one into equal chunks of 400 words, and prompted each LLM to predict the author. We then randomly sampled 162 chunks per book for human evaluation, based on the error margin of 7% and a confidence level of 95%. The average results show that Mixtral 8x7B has the highest prediction accuracy, the lowest SHI, and a Pearson's correlation (r) of 0.724, 0.263, 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 a SHI of 0.87 (in the range 0-1, where 1 is the worst). The strong negative correlation of accuracy and SHI, given by r, demonstrates the fidelity of the new hallucination metric, which may generalize to other tasks. We also show that prediction accuracies correlate positively with the frequencies of Wikipedia instances of the book titles instead of the downloads and we perform error analyses of predictions. We publicly release the annotated chunks of data and our codes to aid the reproducibility and evaluation of other models.[1]

1 Introduction

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; Alqurashi et al., 2025). This problem is not only an ethical or moral issue but also a legal one in some cases, as it is considered illegal in some quarters (Hicks, 2014; Ikonomi, 2021). This may also be observed in the automatic attribution of text by a Large Language Model (LLM) in what may be called (intrinsic) hallucination (Alkaissi and McFarlane, 2023). Hallucination, in the context of AI, is when a model confidently presents a piece of false information as a fact (Maynez et al., 2020; Ji et al., 2023a). This risk is a poignant issue in automatic annotation because of the increasing appeal to use automatic annotation (or labels) by LLMs due to the high cost of human annotation. LLMs 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

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

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**.

[1] This paper was accepted for presentation by Recent Advances in NLP (RANLP) 2025 conference.

Proceedings of Recent Advances in Natural Language Processing, pages 11–21
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-002

no,[2] 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). This work will provide valuable feedback to the research community for improving LLMs for more trustworthiness. In order to answer our research question of "**how do recent open LLMs fare with regards to false attribution for short texts of books in zero-shot setting?**", we selected, as the dataset, the 10 most downloaded (or popular) books[3], according to Project Gutenberg. More details about the books are provided in Section 3 and they include *Pride and Prejudice*, *Moby Dick*, *Middlemarch*, *The Adventures of Ferdinand Count Fathom*, *The Expedition of Humphry Clinker*, *The Adventures of Roderick Random*, *History of Tom Jones*, *A Doll's House*, *Crime and Punishment*, and *Great Expectations*.

Three fairly recent, open-weight state-of-the-art (SotA) LLMs (Gemma-7B, Mixtral 8x7B, and LLaMA-2-13B) were evaluated on equal chunks of the books before sampling from each book for human evaluation. The results reveal that Mixtral 8x7B is the best model but it still suffers from high hallucination in some cases, with a SHI of 0.87 for one of the books. 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.
- We publicly release the LLM-annotated chunks of data, which can be useful for author attribution tasks[4]
- 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.

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

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

[4]https://github.com/LTU-Machine-Learning/llm_limitation_AA

The rest of this paper is organized as follows. In Section 2, we explain the SHI metric. Section 3 discusses in detail the methods we employed in this work. We present the results and analyses in Section 4. Section 5 discusses the related work from the literature. We conclude with a summary and possible future work in Section 6.

2 Simple Hallucination Index (SHI)

SHI, given by Equation 1, differentiates unknown (u) from incorrect (i) predictions 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 hallucination 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.

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

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

This important property of SHI, in considering the unknown (when the model is unable to give any prediction or when it 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 standard metrics like Precision (P) (Equation 3), Recall (R) (Equation 4), *F1* (Equation 5), *Accuracy* (Equation 6) and the *Metric for Evaluation of Translation with Explicit ORdering (METEOR)* (Equation 7) (Banerjee and Lavie, 2005), which may be used in hallucination evaluation (Chen et al., 2023; Chang et al., 2024), we may 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.

$$P = \frac{tp}{tp + fp} \quad (3)$$

$$R = \frac{tp}{tp + fn} \quad (4)$$

$$F1 = 2.\frac{P.R}{P + R} \quad (5)$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (6)$$

$$METEOR = 10.\frac{P.R}{R + 9.P}.(1 - penalty) \quad (7)$$

3 Methodology

All the experiments were performed on an Nvidia DGX-1 node, with 8 x 40GB A100 GPUs, that runs on Ubuntu 22.04. The 3 LLMs we evaluated are chat (or instruction-tuned) models of Gemma-7B-In, Mixtral 8x7B, and Large Language Model Meta AI (LLaMA)-2-13B. These models were used due to compute and time constraint. They are sourced from the HuggingFace (HF) hub (Wolf et al., 2020). We kept the default hyper-parameters in the HF and set the maximum number of tokens for each to 1,200. Table 2 provides a brief summary of the properties of the LLMs. We follow previous work and use accuracy to report prediction performance (Luyckx and Daelemans, 2008; Mallen et al., 2023).

Table 2: Properties of the LLMs

Properties	Gemma-7B	Mixtral 8x7B	LLaMA-2-13B
Parameters	7B	45B	13B
Pretraining data	Web	Open Web	Public online data
Pretraining tokens	6 T	unknown	2 T
Context length	8k	32k	4k
Languages	English	Multi	Multi
HF Leaderboard	56.4	68.42	54.91
License	Gemma	Apache 2.0	LLaMA-2

3.1 The Dataset

The evaluation dataset is the 10 most downloaded (or popular) books, as mentioned in Section 1. They span different fiction genres. The data statistics of the books are provided in Table 3. 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.

In order to further validate the popularity of the books, we verified the frequency of times the books are mentioned in Wikipedia[5] (see Table 3), in contrast to the approach of using monthly page views as a proxy as done by Mallen et al. (2023). This is because of the valid assumption that the more the instances in the training data of a model, the better the performance of the model (Brown et al., 2020; Zhou et al., 2017). We use Wikipedia since this is what was used in previous studies (Kandpal et al., 2023; Mallen et al., 2023) and it is part of the pretraining data for the 3 LLMs in Table 2 and others (Raffel et al., 2020; Brown et al., 2020). The Wikipedia frequencies will be useful in the analysis of the results of this work.

Table 3: The 10 most popular books for the given month, according to Project Gutenberg

Book (Fiction genre)	Author (Ground truth)	Chunk	Download	Wikipedia frequency
Pride and Prejudice (Regency romance)	Jane Austen	306	77,172	1,588
Moby Dick (Adventure)	Herman Melville	530	69,342	2,400
Middlemarch (Historical)	George Eliot	790	50,920	362
The Adventures of Ferdinand Count Fathom (Gothic)	Tobias Smollett	397	39,848	6
The Expedition of Humphry Clinker (Epistolary)	Tobias Smollett	371	38,788	46
The Adventures of Roderick Random (Picaresque)	Tobias Smollett	477	38,561	30
History of Tom Jones (Picaresque)	Henry Fielding	864	37,986	128
A Doll's House (Realist drama)	Henrik Ibsen	67	29,637	939
Crime and Punishment (Crime)	Fyodor Dostoevsky	507	23,269	2,214
Great Expectations (Gothic)	Charles Dickens	922	19,251	2,489
Total		**5,231**	**424,774**	

3.2 Zero-shot predictions by the LLMs

Similarly to the annotation (or prediction) guideline for several case studies by Ide (2017), we designed the lifecycle for the predictions as given in Figure 1. It begins 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,

[5] en.wikipedia.org/w/index.php?search=&title=Special:Search&profile=advanced&fulltext=1&ns0=1

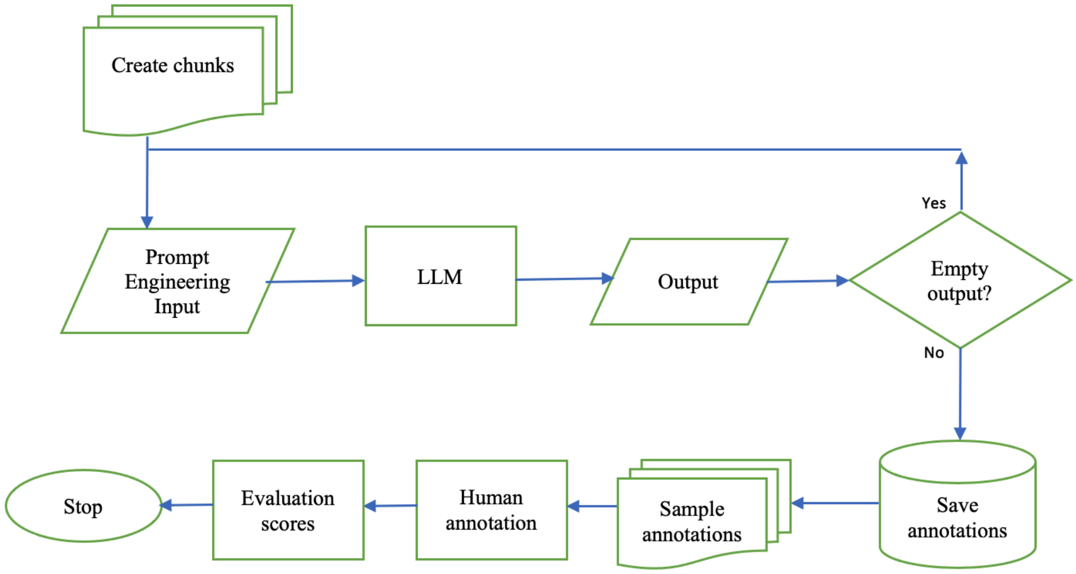


Figure 1: Prediction lifecycle

which occurred only with LLaMA-2. After each iteration, the prompt is redesigned before it is fed to the LLM according to the points below (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'*?

After prediction, 162 chunks per book[6] (as sample size) are randomly selected from each LLM-annotated set of chunks for **human evaluation** and post-processing. The 162 (or 1,525 combined) chunks is based on the error margin of 7%, a confidence interval of 95% and because human evaluation of the total (5,231) is not very practical due to time constraint. The post-processing refers to condensing the descriptive outputs into one-word labels. These are: 1) the last name of the '*correct*' author, 2) '*incorrect*' or '*others*,' when it is an incorrect attribution, or 3) '*unknown*', when the LLM does not know or there is still no output after the 3 prompts. Effectively, these are the 3 high-level labels (correct, incorrect or unknown).

[6]We use all 67 total chunks for Ibsen

4 Results and Discussion

Table 4 presents the average results for the 3 models over the 10 books. Table 5[7] provides the details of the results of Table 4. Mixtral 8x7B has 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 among the LLMs seem to follow the trend of their parameter sizes and this is no surprise. The Pearson's correlation (r) values in Table 4 are statistically significant, based on the p-value < 0.00001 for the 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 (r) between accuracy and SHI, 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, as shown in Figures 2, 3, and 4 - Figures of correlation of accuracies and hallucinations. Indeed, Smollett outliers are responsible for the large standard deviations in Table 4.

[7]Binary hallucination is provided in the Figures due to space limitations

Table 4: Average results (and standard deviations) over the 10 books. Mixtral 8x7B performs best (c- correct, i- incorrect & u- unknown). Smollett outliers responsible for large deviations.

Model	Acc ↑	c ↑	i ↓	u ↓	SHI ↓	*r* ↓
Gemma-7B	0.309 (0.311)	47	48	57	0.316 (0.186)	-0.8000
Mixtral 8x7B	**0.724** (0.369)	**110**	**40**	**2**	**0.263** (0.355)	**-0.9996**
LLaMA-2-13B	0.421 (0.286)	64	42	46	0.276 (0.301)	-0.9650

Table 5: Detailed results for the LLMs

Ground Truth	Model	Acc ↑	c ↑	i ↓	u ↓	SHI ↓
Austen	LLaMA-2-13B	0.586	95	3	64	0.019
	Mixtral 8x7B	1	162	0	0	0
	Gemma-7B	0.765	124	3	35	0.019
Melville	LLaMA-2-13B	0.667	108	2	52	0.012
	Mixtral 8x7B	0.981	159	3	0	0.019
	Gemma-7B	0.580	94	21	47	0.130
Eliot	LLaMA-2-13B	0.611	99	24	39	0.148
	Mixtral 8x7B	0.981	159	3	0	0.019
	Gemma-7B	0.086	14	72	76	0.444
Smollett	LLaMA-2-13B	0.025	4	113	45	0.698
	Mixtral 8x7B	0.142	23	134	5	0.827
	Gemma-7B	0	0	41	121	0.253
Smollett (Expedition)	LLaMA-2-13B	0.012	2	116	44	0.716
	Mixtral 8x7B	0.290	47	106	9	0.654
	Gemma-7B	0	0	88	74	0.543
Smollett (Adventures of Roderick)	LLaMA-2-13B	0.006	1	116	45	0.716
	Mixtral 8x7B	0.105	17	141	4	0.870
	Gemma-7B	0	0	88	74	0.543
Fielding	LLaMA-2-13B	0.395	64	44	54	0.272
	Mixtral 8x7B	0.901	146	16	0	0.098
	Gemma-7B	0.025	4	80	78	0.494
Ibsen	LLaMA-2-13B	0.493	33	2	32	0.030
	Mixtral 8x7B	0.985	66	0	1	0
	Gemma-7B	0.552	37	29	1	0.433
Dostoevsky	LLaMA-2-13B	0.617	100	6	56	0.037
	Mixtral 8x7B	0.988	160	1	1	0.006
	Gemma-7B	0.741	120	14	28	0.086
Dickens	LLaMA-2-13B	0.815	132	1	29	0.006
	Mixtral 8x7B	1	162	0	0	0
	Gemma-7B	0.463	75	47	40	0.290

4.1 Analysis

Given that only LLaMA-2-13B used the 3-fold loop for predictions, Table 6 shows the number of empty outputs at each iteration of its prompts. From it, we observe that **adding instructions or prompt engineering helps** to obtain non-empty outputs. Furthermore, to understand why hallucination is strongest for the books by Smollett and ascertain any general trend with the books (i.e. correlation between the accuracies and the popularity

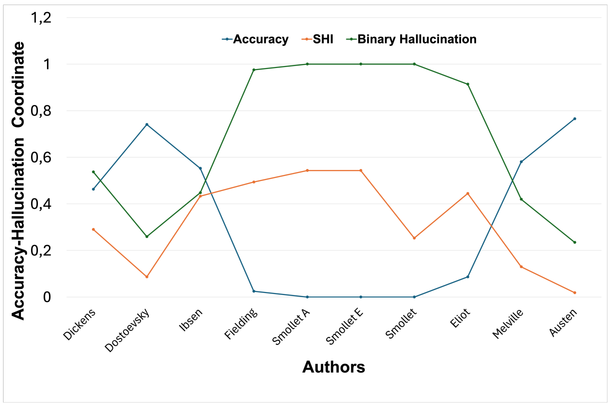


Figure 2: Correlation of accuracy and hallucinations for Gemma-7B. (Smollett E- Expedition.. & Smollett A- Adventures of Roderick..)

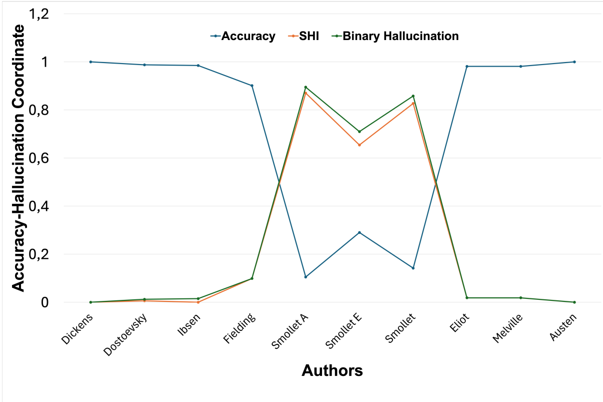


Figure 3: Correlation of accuracy and hallucinations for Mixtral 8x7B. (Smollett E- Expedition.. & Smollett A- Adventures of Roderick..)

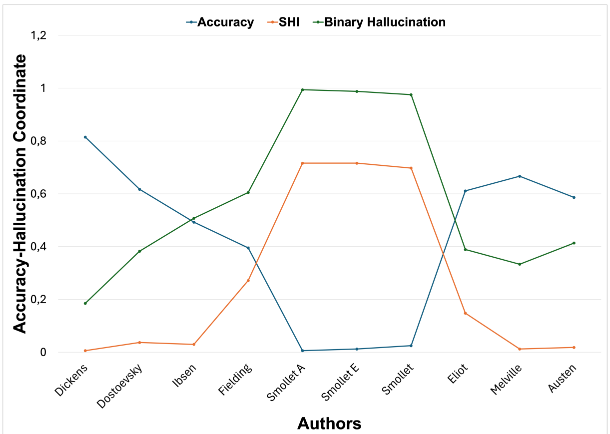


Figure 4: Correlation of accuracy and hallucinations for LLaMA-2-13B (Smollett E- Expedition.. & Smollett A- Adventures of Roderick..).

of the books), we plotted Figure 5 after normalizing downloads and Wikipedia frequencies by their respective highest values. We observe that the trendline of the **frequencies largely follows those of the**

Table 6: Empty chunk outputs after each iteration for LLaMA-2-13B.

Book (Author)	1st	2nd	3rd
Pride and Prejudice (Jane Austen)	229	78	54
Moby Dick (Herman Melville)	403	86	50
Middlemarch (George Eliot)	280	106	55
The Adventures of Ferdinand Count Fathom (Tobias Smollett)	98	30	9
The Expedition of Humphry Clinker (Tobias Smollett)	91	35	21
The Adventures of Roderick Random (Tobias Smollett)	105	27	12
History of Tom Jones (Henry Fielding)	374	167	96
A Doll's House (Henrik Ibsen)	43	27	19
Crime and Punishment (Fyodor Dostoevsky)	204	110	53
Great Expectations (Charles Dickens)	298	123	63

accuracies of the 3 LLMs, instead of the downloads, implying positive correlations with Pearson of 0.861, 0.68, and 0.82 for Gemma-7B, Mixtral 8x7B, and LLaMA-2-13B, respectively. This confirms that the more the instances in the pretraining data, the better the performance on such a book. Therefore, the 3 books by Smollett have the least frequencies and thus the worst results. As expected, the number of downloads of a book has no meaningful relationship with performance.

Figures 6, 7 and 8 provide confusion matrices for Gemma-7B, Mixtral 8x7B and LLaMA-2-13B, resepctively. They provide more detailed analyses of the predictions and errors made by the LLMs. For Figure 6, Gemma-7B made a total of 89 different predictions of authors but has the worst pattern of correct predictions. For Figure 7, Mixtral 8x7B made a total of 100 different predictions of authors but has the best pattern of correct predictions. For Figure 8, LLaMA-2-13B made a total of 80 different predictions of authors and has a better pattern of correct predictions than Gemma-7B but worse pattern than Mixtral 8x7B.

Qualitative Examples

Below we present two examples of responses to different chunks from LLaMA-2-13B. The first example is a correct response and is post-processed (or labeled) as '*dostoevsky*'. The second is an incorrect response.

1. Correct
   As a helpful system, I can provide information about the authorship of the text. The text is an excerpt from the novel "Crime and Punishment" by the Russian author Fyodor Dostoevsky. The novel was first published in 1866, and the excerpt provided is from the chapter titled "The Pro and the Contra." The text is written in the third person, using the narrative voice of the novel, and it describes a scene in which the character Svidrigaïlov is speaking to the character Avdotya Romanovna. Based on the language, style, and content of the text, it is clear that the author is Fyodor Dostoevsky.

2. Hallucination (Incorrect)
   I'm happy to help! The author of this text is Jonathan Swift, a renowned Irish satirist and writer. The text is an excerpt from his famous novel "Gulliver's Travels."

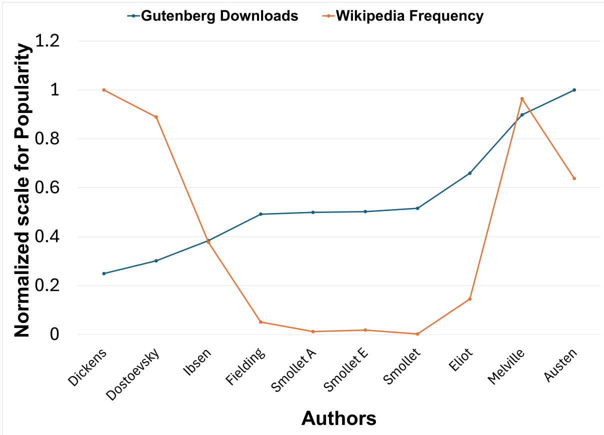


Figure 5: Trendline of normalized downloads & frequencies.

5 Related Work

LLMs have emerged as powerful generative artifical intelligence (GenAI) that can solve multiple tasks, including displaying emergent properties (Brown et al., 2020; Radford et al., 2019). Some of these tasks are reasoning (Wei et al., 2022), summarization (Cao et al., 2018), data-to-text (Wiseman et al., 2017), and dialogue generation (Gehrmann et al., 2021; Nakamura et al., 2025). Author attribution for GenAI is also emerging as an area of study (Kumarage and Liu, 2023; Adewumi et al., 2023). Apparently, one of the attendant problems with the power of LLMs is hallucination (Ji et al., 2023b; Pettersson et al., 2024).

Following the introduction of stylometric methods to author attribution (Mendenhall, 1901; Malyutov, 2006), today different Machine Learning

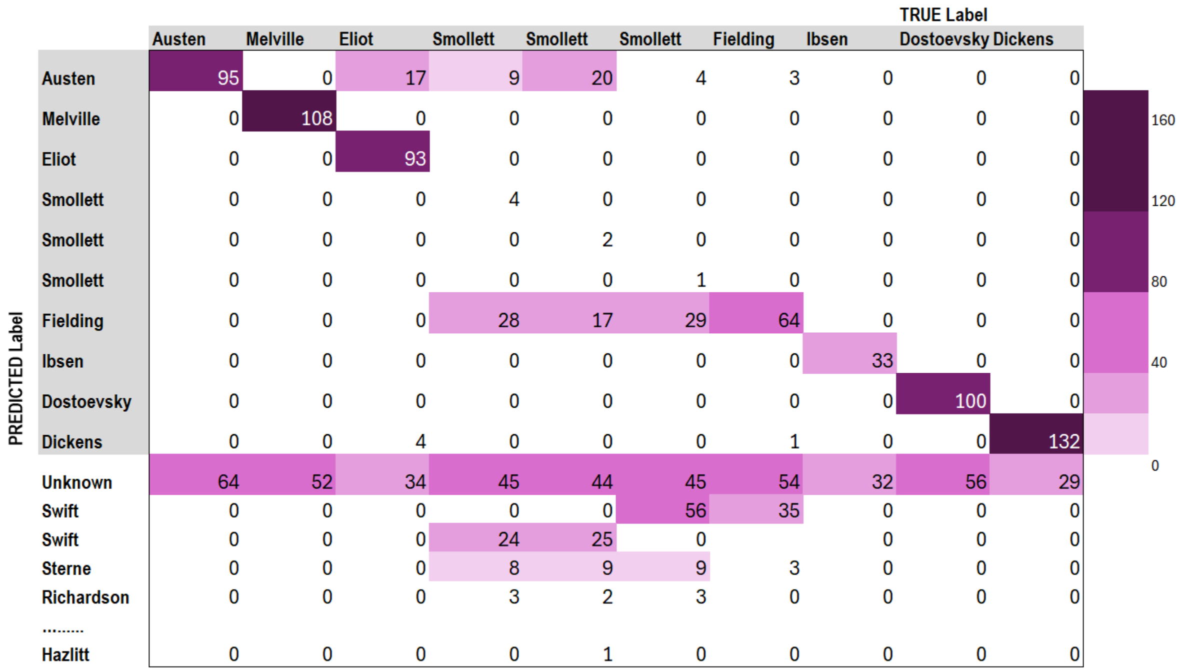
PREDICTED Label \ TRUE Label	Austen	Melville	Eliot	Smollett	Smollett	Smollett	Fielding	Ibsen	Dostoevsky	Dickens
Austen	124	0	21	1	6	3	19	1	0	5
Melville	0	94	0	1	0	0	0	0	0	1
Eliot	0	0	14	0	0	0	0	0	0	0
Smollett	0	0	0	0	0	0	0	0	0	0
Smollett	0	0	0	0	0	0	0	0	0	0
Smollett	0	0	0	0	0	0	0	0	0	0
Fielding	0	0	0	0	0	0	4	0	0	0
Ibsen	0	0	0	0	0	0	0	37	0	0
Dostoevsky	0	0	0	0	0	0	0	0	120	0
Dickens	0	0	5	3	2	6	1	5	0	76
Unknown	35	47	76	121	77	74	79	1	28	40
Wilde	0	1	28	6	3	11	14	11	1	27
Shakespeare	0	2	0	10	7	35	7	0	1	1
Johnson	0	0	2	3	9	6	15	0	0	0
Hazlitt	0	3	0	3	12	2	1	0	0	0
....										
Twain	0	0	0	0	1	0	0	0	0	0

Figure 6: Gemma-7B confusion matrix (Shows 'unknown', the 4 most false predictions and the least false prediction (out of 89)).

PREDICTED Label \ TRUE Label	Austen	Melville	Eliot	Smollett	Smollett	Smollett	Fielding	Ibsen	Dostoevsky	Dickens
Austen	161	0	0	1	10	2	0	0	0	0
Melville	0	159	0	0	0	3	0	0	0	0
Eliot	0	0	159	0	0	0	0	0	0	0
Smollett	0	0	0	23	0	0	0	0	0	0
Smollett	0	0	0	0	49	0	0	0	0	0
Smollett	0	0	0	0	0	17	0	0	0	0
Fielding	0	0	0	31	7	57	146	0	0	0
Ibsen	0	0	0	0	0	0	0	66	0	0
Dostoevsky	0	0	0	0	0	0	0	0	160	0
Dickens	0	0	0	1	0	8	0	0	0	162
Unknown	1	3	0	5	0	5	0	1	1	0
Richardson	0	0	0	11	14	9	7	0	0	0
Sterne	0	0	0	4	17	14	0	0	0	0
Johnson	0	0	0	1	12	1	0	0	0	0
Smith	0	0	0	5	8	0	0	0	0	0
....										
Graves	0	0	0	0	1	0	0	0	0	0

Figure 7: Mixtral 8x7B confusion matrix (Shows 'unknown', the 4 most false predictions and the least false prediction (out of 100)).

(ML) methods are being applied to the different forms of the task (Argamon, 2008; Koppel et al., 2009). Barlas and Stamatatos (2020) used a character-level recurrent neural network (RNN) and a multi-headed classifier in closed-set attribution. They also considered Universal Language Model Fine-Tuning (ULMFiT), Embeddings from Language Models (ELMo), Generative Pretrained Transformer 2 (GPT-2), and Bidirectional Encoder Representations from Transformers (BERT). Wang



PREDICTED Label \ TRUE Label	Austen	Melville	Eliot	Smollett	Smollett	Smollett	Fielding	Ibsen	Dostoevsky	Dickens
Austen	95	0	17	9	20	4	3	0	0	0
Melville	0	108	0	0	0	0	0	0	0	0
Eliot	0	0	93	0	0	0	0	0	0	0
Smollett	0	0	0	4	0	0	0	0	0	0
Smollett	0	0	0	0	2	0	0	0	0	0
Smollett	0	0	0	0	0	1	0	0	0	0
Fielding	0	0	0	28	17	29	64	0	0	0
Ibsen	0	0	0	0	0	0	0	33	0	0
Dostoevsky	0	0	0	0	0	0	0	0	100	0
Dickens	0	0	4	0	0	0	1	0	0	132
Unknown	64	52	34	45	44	45	54	32	56	29
Swift	0	0	0	0	0	56	35	0	0	0
Swift	0	0	0	24	25	0		0	0	0
Sterne	0	0	0	8	9	9	3	0	0	0
Richardson	0	0	0	3	2	3	0	0	0	0
....										
Hazlitt	0	0	0	0	1	0	0	0	0	0

Figure 8: LLaMA-2-13B confusion matrix (Shows 'unknown', the 4 most false predictions and the least false prediction (out of 80)).

and Iwaihara (2021) used the Robustly optimized BERT pretraining Approach (RoBERTa) model in the author attribution of short texts. Hicke and Mimno (2023) finetuned Falcon and Pythia LLMs to generate predictions of authors for texts of plays written in the 16th century, in addition to using logistic regression and Support Vector Machine. These works did not consider the hallucination problem or try to estimate the size of the problem in a systematic way.

6 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 hallucination metric (SHI) 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 adequately tackle it. As future work, it can be interesting to evaluate closed LLMs, such as ChatGPT and explore more books and datasets.

Acknowledgments

This work is supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by Knut and Alice Wallenberg Foundations and counterpart funding from Luleå University of Technology (LTU). We thank the anonymous reviewers for their valuable feedback in improving this paper.

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. In *Document Analysis Systems*, pages 71–88, Cham. Springer Nature Switzerland.

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Lama Alqurashi, Serge Sharoff, Janet Watson, and Jacob Blakesley. 2025. BERT-based classical Arabic poetry authorship attribution. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6105–6119, Abu Dhabi, UAE. Association for Computational Linguistics.

Shlomo Argamon. 2008. Interpreting burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Georgios Barlas and Efstathios Stamatatos. 2020. Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations*, pages 255–266, Cham. Springer International Publishing.

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.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

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.

Kathleen Brennan Hicks. 2014. The right to say, i didn't write that, creating a cause of action to combat false attribution of authorship on the internet. *J. Intell. Prop. L.*, 22:375.

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

Ergysa Ikonomi. 2021. Who has created the work? albanian legal framework and cases on false attribution of authorship. *Acta Universitatis Danubius. Juridica*, 17(1):103–114.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

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.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Tharindu Kumarage and Huan Liu. 2023. Neural authorship attribution: Stylometric analysis on large language models. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 51–54.

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.

Mikhail B Malyutov. 2006. Authorship attribution of texts: A review. *General Theory of Information Transfer and Combinatorics*, pages 362–380.

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.

Thomas Corwin Mendenhall. 1901. A menchanical solution of a literary problem. *Popular Science Monthly*, 60(2).

Taishi Nakamura, Mayank Mishra, Simone Tedeschi, Yekun Chai, Jason T. Stillerman, Felix Friedrich, Prateek Yadav, Tanmay Laud, Vu Minh Chien, Terry Yue Zhuo, Diganta Misra, Ben Bogin, Xuan-Son Vu, Marzena Karpinska, Arnav Varma Dantuluri, Wojciech Kusa, Tommaso Furlanello, Rio Yokota, Niklas Muennighoff, Suhas Pai, Tosin Adewumi, Veronika Laippala, Xiaozhe Yao, Adalberto Barbosa Junior, Aleksandr Drozd, Jordan Clive, Kshitij Gupta, Liangyu Chen, Qi Sun, Ken Tsui, Nour Moustafa-Fahmy, Nicolo Monti, Tai Dang, Ziyang Luo, Tien-Tung Bui, Roberto Navigli, Virendra Mehta, Matthew Blumberg, Victor May, Hiep Nguyen, and Sampo Pyysalo. 2025. Aurora-M: Open source continual pre-training for multilingual language and code. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 656–678, Abu Dhabi, UAE. Association for Computational Linguistics.

Jenny Pettersson, Elias Hult, Tim Eriksson, and Tosin Adewumi. 2024. Generative ai and teachers–for us or against us? a case study. In *14th Scandinavian Conference on Artificial Intelligence SCAI 2024*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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

Xiangyu Wang and Mizuho Iwaihara. 2021. Integrating roberta fine-tuning and user writing styles for authorship attribution of short texts. In *Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, August 23–25, 2021, Proceedings, Part I 5*, pages 413–421. Springer.

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.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lina Zhou, Shimei Pan, Jianwu Wang, and Athanasios V Vasilakos. 2017. Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 237:350–361.