Quantifying Contextual Hallucinations in NLP Research Papers Before and After the LLM Era

Adiba Ibnat Hossain*, Miftahul Jannat Mokarrama and Hamed Alhoori

Northern Illinois University, DeKalb, IL, USA

Abstract

The emergence of Large Language Models (LLMs) has raised growing attention to their potential implications for academic writing. Although LLMs demonstrate impressive generative capabilities, they are also known to produce inaccurate yet confidently presented content, a phenomenon known as hallucination. When such hallucinated content appears in scientific writing, it can undermine clarity, introduce context inconsistencies, and potentially compromise the integrity of the research narrative. In this study, we focus on detecting one specific type of hallucination, namely “context inconsistency", a form of faithfulness hallucination. We investigate the prevalence of contextual contradictions in NLP research papers across two distinct periods: before and after the widespread availability of LLMs, treating NLP as a case study. The paper explores the evolution of inconsistencies, identifies potential LLM-induced discrepancies in the post-LLM era, and evaluates their severity at the paragraph level. We conducted an in-depth analysis of the frequency and intensity of contextual hallucination and observed an increase in inconsistent research articles over time, especially in 2023 and 2024. Interestingly, while inconsistency rates rise overall, the share attributed to AI-generated text has declined, hinting that such content is becoming harder to distinguish from human writing. Dataset & Code

Keywords

Hallucination, LLM, Scientific Writing, Context Inconsistency, Faithfulness, NLP, Academic Integrity

1. Introduction

Maintaining scientific integrity is essential in research communication. With the increasing use of LLMs in generating scientific text, new challenges can arise, such as hallucinated content. When AI-generated content appears non-sensical or unfaithful to the source content, it is identified as hallucination [1]. LLMs have revolutionized natural language processing, but their tendency to hallucinate raises concerns about their reliability in real-world applications [2]. The trustworthiness of LLMs as scientific writing assistants is still questionable, as they exhibit several issues like introducing contextual drift [3], mixing authentic and fabricated facts [4], exaggerating scientific findings [5], and a high hallucination rate in generating a literature review [6]. A study [7] raises the concern that irresponsible use of LLM for the generation and dissemination of scientific articles can lead to adverse consequences. While studies on hallucination detection are quite popular for tasks like question-answering [8], open-ended summarization [9], dialogue systems [10], and machine translation [11], there is still a gap in exploring the prevalence of hallucination in research papers. Such risk impacts digital libraries, which preserve human knowledge while providing essential infrastructure for research and discovery [12, 13]. Hence, for tasks like writing research papers, where the truth matters, detecting hallucinations is crucial [14]. A notable step in this direction is SciHal [15], a hallucination detection dataset for scientific content where experts identify hallucinated statements provided by GenAI-powered research assistants.

The scope of hallucination is conceptually broad, and a survey [16] addresses this by redefining the taxonomy of hallucinations exclusively for LLM applications. It categorizes hallucinations into two major types: 1) *factuality hallucination* and 2) *faithfulness hallucination* [17]. The latter refers to

AI4SciSci'25: Workshop on the Artificial Intelligence and the Science of Science, December 15, 2025, DeKalb, Illinois, USA

*Corresponding author.

ahossain4@niu.edu (A. I. Hossain); mmokarrama@niu.edu (M. J. Mokarrama); alhoori@niu.edu (H. Alhoori)
https://adiba1604041.github.io/portfolio/ (A. I. Hossain); https://jannatmokarrama07.github.io/portfolio/ (M. J. Mokarrama); https://alhoori.github.io/ (H. Alhoori)
0009-0007-1280-1584 (A. I. Hossain); 0000-0002-3239-2956 (M. J. Mokarrama); 0000-0002-4733-6586 (H. Alhoori)

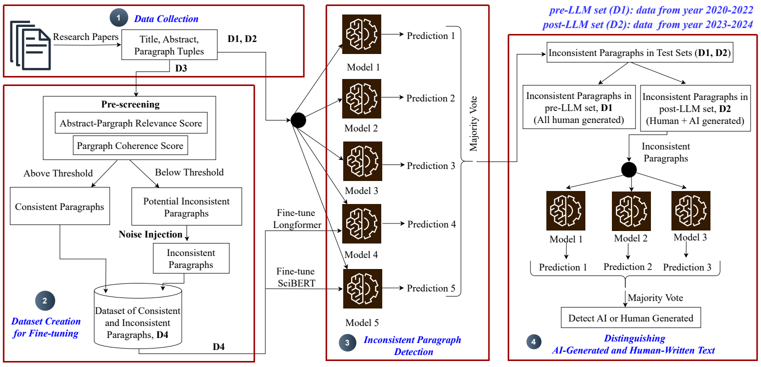

1
Data Collection
Research Papers
Title, Abstract, Paragraph Tuples
D1, D2
D3
Pre-screening
Abstract-Pargraph Relevance Score
Pargraph Coherence Score
Above Threshold
Below Threshold
Consistent Paragraphs
Potential Inconsistent Paragraphs
Noise Injection
Inconsistent Paragraphs
Dataset of Consistent and Inconsistent Paragraphs, D4
D4
2
Dataset Creation for Fine-tuning
Fine-tune Longformer
Fine-tune SciBERT
Model 1
Model 2
Model 3
Model 4
Model 5
Prediction 1
Prediction 2
Prediction 3
Prediction 4
Prediction 5
Majority Vote
3
Inconsistent Paragraph Detection
pre-LLM set (D1): data from year 2020-2022
post-LLM set (D2): data from year 2023-2024
Inconsistent Paragraphs in Test Sets (D1, D2)
Inconsistent Paragraphs in pre-LLM set, D1 (All human generated)
Inconsistent Paragraphs in post-LLM set, D2 (Human + AI generated)
Inconsistent Paragraphs
Model 1
Model 2
Model 3
Prediction 1
Prediction 2
Prediction 3
Majority Vote
Detect AI or Human Generated
4
Distinguishing AI-Generated and Human-Written Text

Figure 1: Overview of contextual hallucination quantification in NLP research papers.

the divergence of generated material from user input or the absence of self-consistency in the created content. To detect hallucination in research papers, our study aims to concentrate on **context inconsistency** [18], a subcategory of faithfulness hallucination. We perform a quantitative analysis of contextual hallucination in research papers to explore the following research questions: **RQ1)** Has the frequency of research papers with inconsistencies changed between the pre-LLM and post-LLM eras? **RQ2)** In the post-LLM era, how prevalent are AI-induced inconsistent research papers? **RQ3)** Are human-created inconsistencies outweighed by AI-generated discrepancies in the post-LLM era? **RQ4)** How significantly does hallucination affect research papers on the paragraph level?

2. Methodology

In our work, we defined 2020–2022 as the pre-LLM era and 2023–2024 as the post-LLM era, based on the broader availability and adoption of LLMs following the release of ChatGPT in late 2022 [19]. Inconsistencies in pre-LLM papers are assumed to be human-induced, while post-LLM discrepancies may stem from humans or LLMs. Focusing on NLP papers, we quantify contextual hallucination at the paragraph level, as surrounding text is required to understand the context. The following sub-sections go into greater detail about the workflow mentioned in Figure 1.

(A) Data Collection

We collected research papers published between 2020 and 2024 from the ACL conference proceedings, extracted title-abstract-paragraph tuples, and organized them into two test sets: **D1** for the pre-LLM era, and **D2** for the post-LLM era (Table 1). In addition, we sampled 50 ACL papers from 2020 to create dataset **D3**, which served as the foundation for the fine-tuning dataset used in the next step.

Table 1
Yearly distribution of collected research papers and extracted title-abstract-paragraph tuples.

Timeframe	Year	Number of Papers	Tuples	Dataset ID
Pre-LLM	2020	399	5777	D1
	2021	399	6764	
	2022	400	7005	
Post-LLM	2023	399	8691	D2
	2024	397	8185	
Total	-	1994	36422	-

(B) Building the Fine-tuning Dataset

To detect inconsistent paragraphs, we employed a set of LLMs as judges and compared their predictions with human annotations. Fine-tuning these models required a labeled dataset, one that distinguishes between consistent and inconsistent text. We defined a paragraph as *consistent* if it was both internally coherent and aligned in meaning with its corresponding abstract. In contrast, a paragraph was identified as *inconsistent* if it lacked coherence or included sentences that diverged from the content or claims made in the abstract. Manually labeling large volumes of papers would have been time-consuming, so we adopted a semi-automated strategy involving two steps: (1) **pre-screening**, which filtered high-quality candidates for consistent and inconsistent samples using a pre-trained sentence transformer [20], and (2) **noise injection**, which introduced irrelevant or misleading sentences into paragraphs to enrich the inconsistent class. During pre-screening, two metrics were computed to guide selection:

(i) **Abstract–Paragraph Relevance Score**: This metric measures how closely a paragraph relates to its abstract. It is calculated as a weighted combination of three criteria: (1) *Contextual Alignment*, which evaluates whether the paragraph supports the main message of the abstract; (2) *Topical Relevance*, which considers whether the content is related to the research focus described in the abstract; and (3) *Purpose Fit*, which examines if the paragraph expands on the purpose outlined in the abstract.

(ii) **Paragraph Coherence Score**: This metric captures the internal coherence of a paragraph. It is calculated as a weighted combination of three criteria: (1) *Contextual Consistency*, which evaluates whether each sentence aligns with its adjacent sentences; (2) *Logical Flow*, which considers whether the paragraph follows a coherent progression; and (3) *Relevance*, which assesses whether all sentences contribute meaningfully to the paragraph's main idea. Guided by exploratory data analysis, we set thresholds above the 60th percentile (0.69 for relevance and 0.68 for coherence) to balance sufficient sample size with high semantic quality. Higher-scoring samples are assumed to be more likely consistent, while tuples falling below both thresholds were marked as "potentially inconsistent".

To make the inconsistent samples reflect genuine divergence, we injected irrelevant sentences into the lower-scoring paragraphs. These sentences were generated using the Mistral-7B-Instruct model [21], guided by prompts designed to produce off-topic or misleading content, and were inserted at random positions within the original paragraphs. The final dataset (**D4**) comprised 450 labeled tuples, each containing a title, abstract, paragraph, and binary label. To mirror realistic imbalances in academic writing quality, we preserved a 61:39 ratio of inconsistent to consistent samples.

(C) Inconsistent Paragraph Detection

To detect inconsistencies in real-world research articles, we evaluated whether paragraphs in D1 and D2 were consistent with their corresponding abstracts, treating each abstract-paragraph pair as a classification instance. We experimented with two pretrained language models, Longformer [22] and SciBERT [23], under five configurations: (a) fine-tuned Longformer, (b) few-shot Longformer, (c) zero-shot Longformer, (d) fine-tuned SciBERT, and (e) zero-shot SciBERT. Longformer is designed for efficient processing of long sequences through sparse attention mechanisms, whereas SciBERT is pre-trained on scientific corpora and thus well-suited for domain-specific tasks. Fine-tuning was carried out using the labeled dataset D4 (Table 2). Each configuration independently predicted whether a given abstract-paragraph pair was consistent or inconsistent. A final label was assigned to each instance through majority voting across the five configurations. On average, 71% of the model configurations agreed on the assigned label, indicating moderate consensus among the setups. To assess the ensemble's accuracy, we compared the majority-voted predictions against human annotations on a representative subset of 100 abstract–paragraph pairs. This evaluation yielded an overall accuracy of 85%, showing that the ensemble's labels aligned with human judgments in most cases.

(D) Paragraph Classification: Distinguishing AI-Generated and Human-Written Text

To investigate the origin of inconsistencies in post-LLM papers, we filtered out all abstract-paragraph tuples labeled as inconsistent from D2. These were then evaluated using three LLMs, acting as judges. The models used were: a) Mistral-7B-Instruct [21], b) Llama-2-7b-chat-hf [24], c) Deepseek-llm-7b-chat [25]. Each model was prompted using a chain-of-thought (CoT) approach guided by five reasoning criteria: (1) *Perplexity*, the degree of predictability in wording; (2) *Burstiness*, the variation in sentence length, structure, and rhythm; (3) *Personalization*, the presence of personal references or subjective voice; (4) *Logical Flow*, the natural progression of ideas across sentences; and (5) *Imperfections*, the occurrence of natural flaws or informal language. Each criterion contributed up to 1 point. Paragraphs with an AI-like score ≥ 3 were labeled as "AI-generated"; otherwise, they were labeled as "Human-written." Final labels were determined through majority voting across the three models, with an average consensus of 81%. To validate this automated classification, we evaluated 100 inconsistent paragraphs using the same criteria outlined in the prompts. The comparison between human annotations and model ensemble outputs yielded an accuracy of 82%, indicating reasonable alignment between model predictions and human judgment.

3. Result and Analysis

(A) RQ1: Inconsistency Over Time: To investigate changes in the frequency of inconsistent papers over time, we employed two evaluation methods: binary judgment and a 5-point Likert scale. In the binary setup, a paper was classified as inconsistent if it contained at least one paragraph labeled inconsistent. This analysis shows an upward trend in inconsistent papers over the years, with a notable spike occurring between 2022 and 2023 (Figure 2a). To gain a more nuanced view, we also conducted a 5-point Likert scale analysis, which categorizes papers based on the percentage of inconsistent paragraphs they contain. The categories were defined as follows: not at all inconsistent (0% - 20%), not very inconsistent (21% - 40%), somewhat inconsistent (41% - 60%), very inconsistent (61% - 80%), extremely inconsistent (81% - 100%). This finer-grained evaluation reinforces the earlier observation: the pre-LLM era (2020–2022) exhibited relatively fewer inconsistencies compared to the post-LLM era (2023–2024) (Figure 2b). Overall, the results suggest that the rise of LLMs coincides with an increase in contextually inconsistent content in academic papers.

(B) RQ2: Post-LLM Hallucination: Following the observed rise in contextual inconsistencies, we analyzed post-LLM papers to estimate what proportion of these research papers could be attributed to AI-driven hallucinations. Between 2023 and 2024, this proportion declined from 86.22% to 75.82%. The trend suggests that although the overall number of papers containing inconsistencies continues to increase, the relative share of inconsistencies attributable to AI is decreasing. One possible explanation is that hallucinated content generated by LLMs has become more subtle, making it harder to distinguish from human-written text.

(C) RQ3: Human vs. AI Errors: To determine whether inconsistencies in the post-LLM research papers are more likely to originate from humans or AI, we examined inconsistent paragraphs from the post-LLM period. A paper was counted as AI-influenced if it contained at least one AI-generated inconsistent paragraph; otherwise, it was considered human-originated. Figure 3a shows that AI-induced inconsistencies were more common than human-originated ones, although their share declined in 2024 compared to 2023.

(D) RQ4: Hallucination Severity: Here we examine the **intensity of hallucination** at the paragraph level using the Abstract-Paragraph Relevance Score, where greater divergence indicates a stronger likelihood of hallucination. Severity was categorized on a 5-point Likert scale: extremely hallucinated (0.00–0.20), very hallucinated (0.21–0.40), somewhat hallucinated (0.41–0.60), not very hallucinated (0.61–0.80), and not at all hallucinated (0.81–1.00). For analysis, paragraphs with scores of 0.00–0.60 were grouped as intensely hallucinated, while those above 0.60 were considered less intensely hallucinated. As shown in Figure 3b, papers published in 2023 contain a larger share of intensely hallucinated paragraphs than those from 2024. This suggests that while hallucinations remain present, their severity is decreasing. AI-assisted writing may now produce subtler inconsistencies, reducing risks to research integrity, and LLMs appear to be improving in fidelity and reliability, though issues still persist.

4. Conclusion

This study revealed a rising trend in NLP research papers containing contextual inconsistencies over time. Within the post-LLM era, AI-generated inconsistencies outweighed those introduced by humans, though their share showed a decline from 2023 to 2024. This decline suggests that hallucinated text produced by LLMs may be becoming less frequent or increasingly difficult to distinguish from human writing. Notably, research papers from 2023 contained a higher proportion of hallucinated content, both in frequency and severity, while papers from 2024 exhibited a reduction in both measures. These findings highlight that although inconsistencies remain a challenge, the nature of AI-generated text is evolving, pointing to gradual improvements in LLM fidelity but also raising concerns about subtler, harder-to-detect risks to scientific integrity.

5. Declaration on Generative AI

GenAI tools were minimally used for grammatical improvements, linguistic refinement, and prompt phrasing.

References

[1] K. Filippova, Controlled hallucinations: Learning to generate faithfully from noisy data, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 864–870. doi:10.18653/v1/2020.findings-emnlp.76.

[2] G. Hao, J. Wu, Q. Pan, R. Morello, Quantifying the uncertainty of llm hallucination spreading in complex adaptive social networks, Scientific reports 14 (2024) 16375.

[3] S. Liu, K. Halder, Z. Qi, W. Xiao, N. Pappas, P. M. Htut, N. A. John, Y. Benajiba, D. Roth, Towards long context hallucination detection (2025). URL: https://www.amazon.science/publications/towards-long-context-hallucination-detection.

[4] H. Alkaissi, S. I. McFarlane, Artificial hallucinations in chatgpt: implications in scientific writing, Cureus 15 (2023).

[5] Times of India, AI models like ChatGPT and DeepSeek frequently exaggerate scientific findings, study reveals, https://timesofindia.indiatimes.com/technology/tech-news/ai-models-like-chatgpt-and-deepseek-frequently-exaggerate-scientific-findings-study-reveals/articleshow/121189880.cms, Updated May 16, 2025.

[6] M. Chelli, J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, J.-L. Raynier, G. Clowez, P. Boileau, C. Ruetsch-Chelli, et al., Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: comparative analysis, Journal of medical Internet research 26 (2024) e53164.

[7] B. Mittelstadt, S. Wachter, C. Russell, To protect science, we must use llms as zero-shot translators, Nature human behaviour 7 (2023) 1830–1832.

[8] B. Snyder, M. Moisescu, M. B. Zafar, On early detection of hallucinations in factual question answering, in: Proceedings of the 30th ACM SIGKDD, 2024, pp. 2721–2732.

[9] D. Wan, K. Sinha, S. Iyer, A. Celikyilmaz, M. Bansal, R. Pasunuru, Acueval: Fine-grained hallucination evaluation and correction for abstractive summarization, in: Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 10036–10056.

[10] N. Dziri, A. Madotto, O. Zaïane, A. J. Bose, Neural path hunter: Reducing hallucination in dialogue systems via path grounding, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 2197–2214. URL: https://aclanthology.org/2021.emnlp-main.168/. doi:10.18653/v1/2021.emnlp-main.168.

[11] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, A. F. Martins, Hallucinations in large multilingual translation models, Transactions of the Association for Computational Linguistics 11 (2023) 1500–1517.

[12] H. Kroll, C. K. Kreutz, M. Jehn, T. Risse, Requirements for a digital library system: A case study in digital humanities, in: Proceedings of the 24th ACM/IEEE JCDL, 2024, pp. 1–3.

[13] A. Zielinski, S. Hirzel, S. Arnold-Keifer, Enhancing digital libraries with automated definition generation, in: Proceedings of the 24th ACM/IEEE JCDL, 2024, pp. 1–11.

[14] Use of AI is seeping into academic journals—and it's proving difficult to detect, Wired (2023). URL: https://www.wired.com/story/use-of-ai-is-seeping-into-academic-journals-and-its-proving-difficult-to-detect/?utm_source=chatgpt.com.

[15] D. Li, B. Palfi, C. Zhang, J. Subramanian, A. Raudaschl, Y. Kakita, A. De Waard, Z. Afzal, G. Tsatsaronis, Overview of the SciHal25 shared task on hallucination detection for scientific content, in: T. Ghosal, P. Mayr, A. Singh, A. Naik, G. Rehm, D. Freitag, D. Li, S. Schimmler, A. De Waard (Eds.), Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 307–315. URL: https://aclanthology.org/2025.sdp-1.29/. doi:10.18653/v1/2025.sdp-1.29.

[16] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025) 1–55.

[17] Y. Dong, J. Wieting, P. Verga, Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization, in: EMNLP 2022, 2022, pp. 1067–1082. doi:10.18653/v1/2022.findings-emnlp.76.

[18] E. Ainsworth, J. Wycliffe, F. Winslow, Reducing contextual hallucinations in large language models through attention map optimization, Authorea Preprints (2024).

[19] A. Bick, A. Blandin, D. J. Deming, The rapid adoption of generative ai, 2024. URL: https://www.stlouisfed.org/on-the-economy/2024/sep/rapid-adoption-generative-ai.

[20] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[21] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[22] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[23] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: EMNLP-IJCNLP, 2019, pp. 3615–3620. doi:10.18653/v1/D19-1371.

[24] H. Touvron, L. Martin, K. Stone, et al., Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[25] DeepSeek-AI, :, X. Bi, D. Chen, G. Chen, et al., Deepseek llm: Scaling open-source language models with longtermism, 2024. arXiv:2401.02954.

A. Appendix

A.1. Supplementary Results

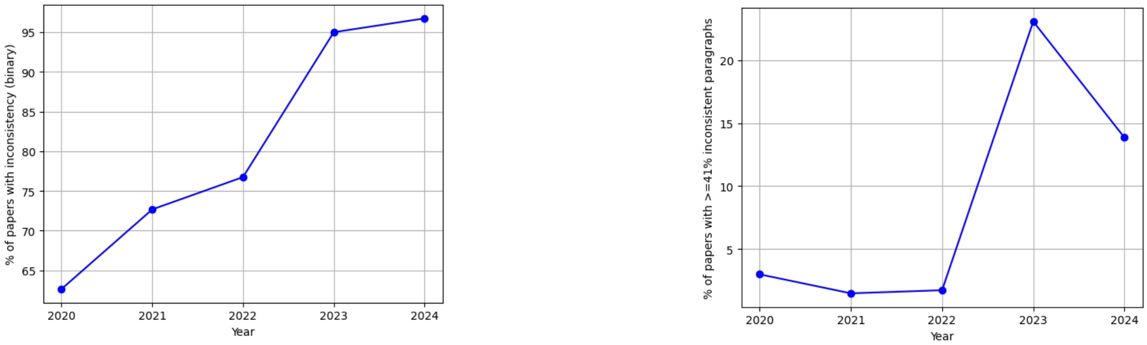
A.1.1. Fine-Tuned Model Performance on Inconsistency Detection

Table 2
Performance of fine-tuned models on inconsistency detection.

	Accuracy	Precision	Recall	F1 score
Longformer	97.78%	97.90%	97.78%	97.79%
SciBERT	97.77%	94.59%	100%	97.22%

A.1.2. Figures Supporting Research Questions

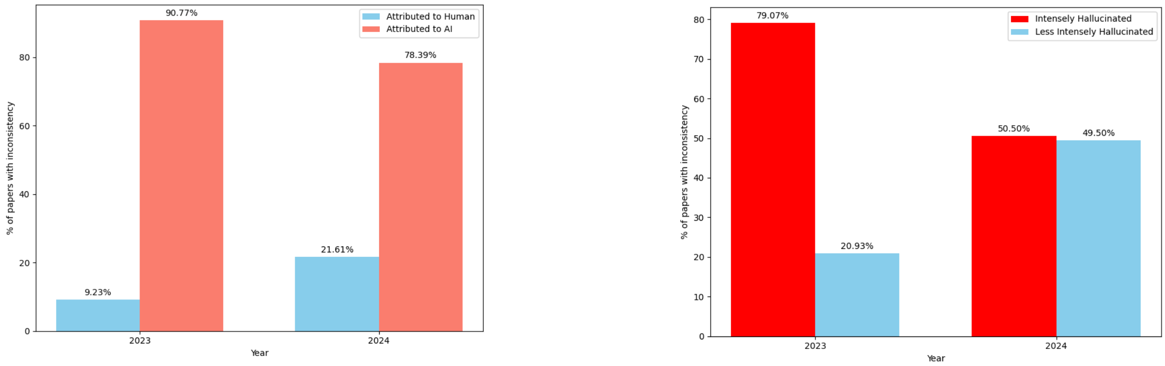
Figures 2a, 2b, 3a, 3b present detailed visualizations supporting the research questions discussed in Section 3.



(a) Binary Judgment

(b) Likert-scale Judgment

Figure 2: (a) Inconsistency trend by year based on binary judgment. (b) Proportion of research papers with $\geq 41\%$ inconsistent paragraphs across publication years.



(a) Human vs. AI Inconsistencies

(b) Hallucination Severity

Figure 3: (a) Comparison of human- and AI-induced inconsistencies (2023–2024). (b) Distribution of less intense and intense hallucinations in research papers published in 2023 and 2024.