Reducing Hallucinations in LLMs: Utilizing STORM for Accuracy in Scientific Writing

1st Jonathan
Computer Science Department
School of Computer Science
Bina Nusantara University
Tangerang, Indonesia 15143
jonathan062@binus.ac.id

2nd Ricky Herald Cenniago
Computer Science Department
School of Computer Science
Bina Nusantara University
Tangerang, Indonesia 15143
ricky.cenniago@binus.ac.id

3rd Steven Chowina
Computer Science Department
School of Computer Science
Bina Nusantara University
Tangerang, Indonesia 15143
steven.chowina@binus.ac.id

***Abstract*—**

***Index Terms*—**

I. Introduction

Artificial Intelligence (AI) has become a key technology with its usage expanding across various fields. AI adoption has surged across industries, with 65% of organizations routinely using AI, nearly doubling from the previous year [2]. A key subfield of AI is Generative AI, which focuses on creating new data outputs, including text, images, audio, and even video, based on learned patterns from large datasets. This technology allows AI models to generate human-like content, making it highly valuable for applications in creative industries, customer service, and scientific writing

A subset of Generative AI is Large Language Models (LLMs), which specialize in natural language understanding and text generation. LLMs, such as GPT-3, GPT-4, Claude, and Gemini, have demonstrated advanced capabilities in producing coherent text, programming code, and research summaries. These models rely on massive datasets to train deep learning architectures, enabling them to respond contextually and accurately to user inputs. The growing adoption of LLMs has transformed various domains, including scientific writing, where they serve as writing assistants for summarization, paraphrasing, and content drafting. However, their reliance on statistical inference rather than true understanding introduces risks, particularly in academic contexts, where precision is critical. Scientific progress in AI is heavily dependent on research papers, as most breakthroughs in AI originate from academic publications. Fundamental AI innovations, such as Gated Recurrent Units (GRUs) in 2014 by Cho et al. [3] and the Transformer architecture in 2017 by Vasnawi et al. [4], were initially proposed in research papers before becoming foundational to modern AI systems. This underscores the importance of academic research in advancing AI technologies.

With the increasing reliance on LLMs in academic writing, there is an opportunity to utilize these models to support research documentation. However, a major challenge is hallucination, where LLMs generate incorrect or fictitious information presented as factual [5]. This issue is particularly concerning in scientific research, where precision and factual accuracy are imperative.

To mitigate this problem, Stanford University developed STORM (Synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking). Unlike conventional LLMs, STORM enhances writing accuracy by integrating retrieval-based information sourcing and multi-agent questioning to refine content. This framework ensures that generated text is factually accurate, properly structured, and supported by valid citations.

The emergence of STORM highlights the significant potential of AI in improving the accuracy of scientific writing. Its effectiveness has been demonstrated through early evaluations, which show that STORM outperforms conventional generative AI approaches. Studies indicate that STORM-generated articles are 25% more structured and 10% broader in scope compared to conventional retrieval-augmented generation methods.

This study aims to analyze and compare STORM with conventional AI-assisted writing methods in reducing hallucinations through a Systematic Literature Review (SLR). The primary objective is to evaluate STORM's effectiveness in producing accurate, well-structured, and reliable scientific content.

The central hypothesis of this research is that STORM outperforms conventional generative AI approaches in terms of factual accuracy, structural organization, and citation reliability. Through comparative analysis, this study seeks to demonstrate STORM's superiority in mitigating hallucinations in scientific writing, while also identifying potential areas for further refinement and development.

II. Literature Review

A. Hallucinations in LLMs and Their Challenges in Scientific Writing

Hallucinations in AI refer to model-generated outputs that sound convincing but are actually incorrect or entirely fictitious. This phenomenon has been documented in various recent studies [5] and has become a serious concern when LLMs are used for tasks that require high accuracy. In scientific writing, hallucinations can take the form of false factual statements or fabricated references.

Huang et al. (2023) noted that hallucinations occur because language models tend to fabricate information when their knowledge is limited or when they fail to find the correct answer, particularly when the model is not connected to external knowledge sources [5]. This issue is exacerbated by LLMs' fluent and natural-sounding language, making incorrect outputs appear credible.

B. Non-STORM Approaches to Reducing Hallucinations in AI Writing

To mitigate hallucinations, various approaches were developed before the introduction of STORM. One of the most common strategies is incorporating external information retrieval into text generation, known as Retrieval-Augmented Generation (RAG).

Lewis et al. (2020) introduced the RAG framework, which integrates document retrieval into language models: the model extracts relevant paragraphs/documents from a knowledge base or the internet, then generates responses based on the retrieved content [6]. This approach improves factuality because the model does not rely solely on its limited internal knowledge but instead leverages up-to-date external information.

RAG has proven effective in question-answering and knowledge-intensive tasks. Many AI-powered search engines have adopted RAG, such as Perplexity and Microsoft Bing. These systems conduct web searches and generate responses with cited references. However, they primarily focus on short-form answers rather than generating structured, long-form writing.

Another challenge is ensuring that the model properly understands and filters retrieved information. Some recent studies attempt to enhance this process through multi-step iteration and interactive refinement.

Jiang et al. (2023) proposed Active Retrieval-Augmented Generation, where the model actively performs multiple rounds of retrieval, evaluating whether additional information is needed before composing the final response [7]. Essentially, the model asks itself: ”Is this information sufficient? What is missing?” If necessary, it retrieves additional data. This iterative process reduces both hallucinations and information saturation, making the model more precise in gathering facts.

Apart from retrieval, another approach to improving LLM accuracy is enforcing citation-based responses. Gao et al. (2023) demonstrated that by training and modifying LLMs to produce text with citations, the reliability and verification of AI-generated content significantly improve [7]. In these models, the AI is programmed to reject questions if it lacks clear references and uncertain claims are flagged, preventing AI from generating baseless responses.

This approaches indicate a trend that generative AI factuality can be enhanced by combining verification mechanisms, external knowledge retrieval, and citation enforcement. However, beyond factual accuracy, another challenge is structuring information into well-organized, long-form academic writing. Most pre-STORM methods are limited to paragraphs or specific Q&A tasks. Writing a full research paper or survey requires topic planning, outline generation and ensuring comprehensive coverage. This is a task that traditionally requires human expertise. Therefore, a more advanced AI framework is needed—one that is not only factually correct but also capable of planning and organizing content like an experienced academic writer.

C. STORM Framework and Its Comparison with Other Methods

To bridge this gap, STORM (Synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking) was introduced by Shao et al. (2024). Unlike previous methods, STORM models the pre-writing phase, mimicking the way human writers plan before drafting long-form articles [1].

STORM operates in two main stages: (1) Research & Outline Creation, the system conducts information retrieval on the given topic, identifying multiple perspectives and subtopics. It then constructs an outline, summarizing key discussion points along with supporting references. (2) Content Generation, based on the outline and verified references, STORM writes a structured academic article, ensuring that each statement includes a citation [1].

STORM's approach is inspired by multi-perspective question asking, where the system simulates a discussion between multiple AI agents, each playing a different role. One agent acts as a “critical writer”, another agent serves as a “subject-matter expert” and then they exchange questions and answers, extracting deeper insights from retrieved sources. For example, in an article about quantum computing, the writer-agent might ask: ”What are the historical developments of quantum computing?”. The expert-agent responds with sourced information, while another agent might ask: ”What are the major controversies in quantum computing?”. This multi-agent dialogue allows STORM to explore diverse viewpoints that might be overlooked in simple retrieval-based approaches [1]. The resulting structured outline ensures that key arguments are comprehensively covered, while irrelevant information is filtered out.

The STORM approach has proven effective in improving the quality of LLM-generated long-form content. According to an evaluation conducted on the FreshWiki dataset (a collection of the latest Wikipedia articles) [1], articles produced by STORM demonstrated significant improvements in terms of organizational structure and content coverage compared to baseline methods. Compared to standard RAG models that directly generate responses, articles produced using STORM were 25% more likely to be rated as well-organized and 10% more likely to cover a broader range of relevant topics [1].

Experienced Wikipedia editors stated that STORM-produced content appeared more engaging, structured, and informative compared to articles written without the multi-agent framework. This finding highlights that the pre-writing process (outline generation) employed by STORM plays a crucial role in enhancing the quality of AI-generated writing.

Despite its significant advantages, several limitations of STORM have been identified in academic literature. In a study conducted by Shao et al. (2024), it was found that the verifiability of STORM-generated content still requires further improvement. Approximately half of STORM-generated articles received lower verifiability scores compared to human-written articles [1].

One major issue stems from source bias transfer, where the bias or narrow perspective present in retrieved sources influences the generated content [1]. For example, if the majority of sources retrieved favor a particular viewpoint, STORM-generated content may unintentionally present an imbalanced perspective on the topic.

Another challenge is the over-association of unrelated facts, where STORM, in its effort to integrate multiple pieces of information, sometimes connects unrelated or weakly relevant details within a single narrative. This issue is similar to information overload, where excessive detail dilutes the clarity and relevance of the final content.

Compared to non-STORM approaches, STORM offers a more comprehensive solution for AI-assisted academic writing. Previous methods, such as conventional RAG or automatic citation enforcement, primarily focus on correcting factual inaccuracies at the sentence level. In contrast, STORM addresses the entire writing process—starting from topic planning to structured content generation.

For scientists and researchers, STORM functions as a literature assistant, capable of summarizing broad topics into well-organized sections within an academic paper. However, human intervention and critical evaluation remain necessary to validate the AI-generated content and refine the final interpretation.

Nevertheless, by combining multi-agent dialogue mechanisms with retrieval-based content generation, STORM sets a new benchmark for AI-assisted scientific writing. Compared to previous LLM methods, STORM demonstrates a higher level of effectiveness in minimizing hallucinations, making it a significant step forward in ensuring reliable, AI-generated academic content.

III. Methodology

This research adopts a comprehensive structured experimental approach aimed at evaluating the effectiveness of the STORM framework (Synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking) in minimizing hallucination in scientific writing generated by Large Language Models (LLMs). The term ”hallucination” in this context refers to the generation of information that appears factual but is actually incorrect, unverifiable, or entirely fabricated. This is a pressing issue in the adoption of AI in academic research, where factual accuracy and trustworthiness are critical.

A. Research Methodology

The research methodology consists of six steps:

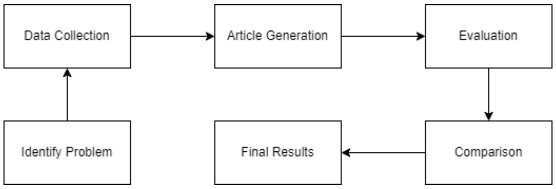

Fig. 1. Research Methodology Flowchart

1) Identify Problem: In the first phase of the research, the main problem is identified. The core issue under investigation is the generation of scientific writing using Large Language Models (LLMs). Current LLM models, especially those not integrated with frameworks like STORM, often produce articles that lack structural coherence, reliable references, and factual accuracy. These weaknesses are seen in the LLM-generated articles without STORM, where there is an increased risk of hallucinations (fabricated information) and weak citations. This motivates the development and evaluation of the STORM framework.

2) Data Collection: Once the problem is identified, the data collection phase begins. For this research, three main data sources are used for comparison:

a) FreshWiki Dataset: A collection of scientifically accurate, human-curated articles from Wikipedia, which serves as the gold standard for reference content.

b) LLM Without STORM: This represents articles generated by an LLM model (e.g., ChatGPT) without using STORM. These articles are expected to contain errors, hallucinations, and structural weaknesses.

c) LLM With STORM: This represents articles generated using the STORM framework, which incorporates structured data retrieval and citation to improve factual accuracy and article structure.

3) Article Generation: After collecting the data, the next step is to generate articles using the selected methodologies. This involves two main processes:

a) FreshWiki: No article generation is required here. The FreshWiki dataset is simply used as a comparison baseline for the evaluation of the AI-generated content.

b) LLM Without STORM: The LLM generates articles based on the given topics without any external framework or citations, leading to a greater chance of hallucinations or inaccuracies.

c) LLM With STORM: The LLM generates articles, but this time, STORM is used to enhance the content. The STORM framework guides the AI by retrieving factual data, generating structured outlines, and adding citations, thus improving accuracy and structural coherence.

4) Evaluation: The Evaluation phase is crucial for assessing the quality of the generated articles across the three approaches: FreshWiki, LLM without STORM, and LLM with STORM. This phase will be conducted using a Mix Method Approach, which combines both quantitative and qualitative assessments.

a) Quantitative Evaluation: The Evaluation phase is crucial for assessing the quality of the generated articles across the three approaches: FreshWiki, LLM without STORM, and LLM with STORM. This phase will be conducted using a Mix Method Approach, which combines both quantitative and qualitative assessments.

b) Qualitative Evaluation: This phase involves gathering subjective feedback from participants (experts or general users) who will review the generated articles and provide insights into the overall quality, readability, and credibility of the content.

5) Comparison: In the Comparison phase, the articles from LLM with STORM, LLM without STORM, and FreshWiki are compared to assess the overall effectiveness of STORM in improving the quality of AI-generated content. Comparison Metrics:

a) Citation Accuracy: Which method generates the most reliable citations

b) Factual Accuracy: Which method provides the most factually correct content

c) Stuctural Coherence: Which method provides the most factually correct content

d) Readibility and User Feedback: Which articles were rated higher by participants in terms of clarity and trustworthiness

6) Final Result: In the Final Result phase, the findings from both the Evaluation and Comparison phases are combined, and conclusions are drawn based on the data collected. This phase summarizes the effectiveness of the STORM framework and its ability to reduce hallucinations, improve article structure, and enhance citation accuracy.

B. Experiment Methodology

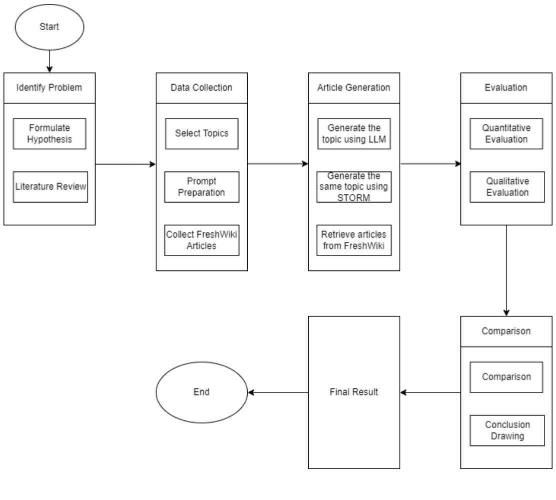

Fig. 2. Experiment Methodology Flowchart

1) Formulate Hypothesis: The research begins by formulating a hypothesis that addresses the central question of the study: Can the STORM framework improve the overall quality of AI-generated academic articles compared to traditional methods? This hypothesis acts as the foundational assumption of the study and forms the basis for the experimental design, evaluation criteria, and expected outcomes. The aim is to explore whether the STORM framework, which integrates retrieval-based generation and structured content creation, can enhance the factual accuracy, structural integrity, and citation reliability of AI-generated articles, thus overcoming the limitations seen in traditional LLM-generated content.

2) Literature Review: A comprehensive literature review follows the hypothesis formulation, providing a theoretical context for the study. In this phase, existing research on Generative AI, Retrieval-Augmented Generation (RAG), and the STORM framework is examined. This step helps to identify gaps in the current literature, especially regarding AI's ability to generate accurate, well-structured, and reliable academic content. The review justifies the need for this study and helps to position STORM as a potential solution for addressing common problems such as hallucinations and poor citation practices in AI-generated writing.

3) Select Topics: Once the problem and theoretical context are established, the next step is to select topics that will serve as the basis for the generated articles. The topics need to be carefully chosen to meet specific criteria: they must be academic in nature, suitable for AI-generated content, and complex enough to test the depth and quality of the generated articles. The chosen topics should also be general enough for the LLMs to handle while still allowing for meaningful evaluations. The goal here is to ensure that the generated content is both relevant and sufficiently challenging to assess the model's ability to handle intricate academic subjects.

4) Prompt Preparation: Once the topics are selected, prompt preparation is a crucial step in ensuring the fairness and consistency of the generated content. For the STORM framework, structured and specific prompts are crafted to guide the model through the process of data retrieval, multi-perspective questioning, and structured outlining. These prompts align with the STORM methodology, which encourages the model to gather information from reliable sources and organize it logically. On the other hand, non-STORM prompts are simpler and more basic, without the framework's structural guidance, ensuring that the comparison between the two models is as fair as possible. This step ensures that both models are evaluated under similar conditions.

5) Collect FreshWiki Articles: At this stage, FreshWiki articles are collected to serve as the benchmark for comparison. These articles are curated and verified by human editors, making them reliable and accurate. FreshWiki articles are used because they are a trusted source of factual information, clarity, and structural coherence. The goal of this phase is to gather a set of articles on the selected topics that will act as a gold standard when comparing them against the AI-generated articles.

6) Generate the Topic Using LLM: The first step in article generation is to create an article for the selected topic using the LLM (Large Language Model) without any additional frameworks. In this process, the model generates the content purely based on the input prompt, relying on its internal knowledge and capabilities. As the model does not use any structured methods, the content may lack proper citations, logical structure, and may be prone to hallucinations (inaccurate or fabricated information). This serves as a baseline to evaluate the quality of AI-generated content.

7) Generate the Same Topic using STORM: Once the article is generated using the LLM alone, the STORM framework is applied to generate a second article on the same topic. The STORM method follows a structured, multi-step approach which includes outline generation, subtopic expansion, data retrieval, and citation integration. This process ensures that the generated article is factually accurate, well-organized, and supported by reliable sources. Applying the STORM framework is expected to improve the article's quality, reduce hallucinations, and provide structured references, making the content more reliable and academic.

8) Retrieve articles from FreshWiki: Alongside the AI-generated content, FreshWiki articles on the selected topics are retrieved for comparison. These articles are used as a benchmark to assess how the generated content matches up against human-curated articles in terms of factual accuracy, citation quality, and structural coherence. FreshWiki serves as the standard that AI-generated content will be compared against in terms of reliability and academic quality.

9) Quantitative Evaluation: After the articles are generated, the next phase is quantitative evaluation. This involves evaluating the generated articles using objective criteria, which are measured using a rubric-based scoring system. The articles are assessed on factors like structural organization, accuracy of citations, factual correctness, depth of coverage, and completeness. This quantitative evaluation provides measurable data that allows for a direct comparison of how well each method (LLM with STORM, LLM without STORM, and FreshWiki) performed according to established academic writing standards. Tools or software may also be used to ensure the evaluations are consistent and reproducible.

10) Qualitative Evaluation: Following the quantitative evaluation, the qualitative evaluation phase gathers subjective feedback from users and experts. This feedback focuses on aspects like readability, credibility, and overall quality. Participants (individuals or groups) who are familiar with AI-generated content provide their opinions on how trustworthy, readable, and well-structured the articles are. This qualitative feedback is crucial in assessing how the articles are perceived by human readers, which is an important aspect of content quality, especially in academic writing.

11) Comparison: Once both the quantitative and qualitative evaluations are completed, the results are compared side by side. The articles generated by LLM with STORM, LLM without STORM, and FreshWiki are assessed to identify the strengths and weaknesses of each method. This comparison helps determine whether STORM improves the quality of the AI-generated articles in terms of accuracy, structure, and citation reliability.

12) Conclusion Drawing: Finally, all the findings from the evaluation and comparison phases are synthesized to draw conclusions. This phase involves analyzing whether STORM significantly improves the quality of AI-generated academic writing. The conclusions will also discuss the implications of STORM for future AI-driven writing models and academic content generation, as well as highlight the limitations of the current study and suggest areas for future research.

IV. Expected Results

This study aims to evaluate the effectiveness of the STORM framework in improving the quality of AI-generated scientific writing. We expect that the integration of STORM will significantly reduce hallucinations and enhance the factual accuracy, structural organization, and citation reliability of the generated content.

We hypothesize that STORM will outperform traditional generative AI models in multiple key areas. First, STORM is expected to produce more factually accurate content by leveraging multi-agent questioning and external data retrieval. This is anticipated to reduce hallucinations, where traditional models tend to fabricate information. Second, STORM's structured outline generation will lead to more logically organized and coherent content compared to models that generate text without a predefined structure. Third, STORM will produce content with higher citation reliability. Unlike traditional models, which often generate text without citations, STORM integrates reliable data sources and ensures proper referencing, improving the credibility of the generated content.

In comparison to Non-STORM models, which typically suffer from hallucinations and poor organization, STORM is expected to provide more comprehensive and well-organized articles. These articles will have a broader scope, ensuring that key aspects of a topic are adequately covered. Furthermore, STORM-generated articles are expected to have reliable citations, addressing one of the major limitations of non-STORM models.

When compared to human-written articles, which generally exhibit high factual accuracy, structural coherence, readability, and citation reliability, we expect STORM-generated content to approach human-level performance in terms of accuracy and structure. However, it is likely that human-written articles will still outperform STORM-generated content in areas such as clarity, readability, and nuanced understanding of complex topics, as humans bring critical thinking and expertise to the writing process.

The evaluation of Non-STORM, STORM, and Human-Written articles will focus on several key metrics: factual accuracy, structural coherence, citation reliability, readability, and coverage of topics. These metrics will provide a comprehensive assessment of the effectiveness of STORM in improving the quality of AI-generated academic writing.

The results of this study are expected to demonstrate that the STORM framework offers a significant improvement over traditional AI models in terms of reducing hallucinations and enhancing the overall quality of generated content. By addressing the challenges of hallucinations and providing a more structured approach to content generation, STORM has the potential to set a new benchmark for AI-assisted academic writing.

V. Result and Discussion

VI. Conclusion

References

[1] Y. Shao et al., “Assisting in Writing Wikipedia-like Articles from Scratch with Large Language Models,” arXiv:2402.14207 [cs.CL], 2024.

[2] McKinsey, ”The state of AI in early 2024: Gen AI adoption spikes and starts to generate value,” 2024.

[3] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] K. Cho et al., “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in Proc. EMNLP, Doha, Qatar, 2014, pp. 1724–1734.

[5] L. Huang et al., “A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions,” arXiv:2308.XXXX, 2023.

[6] P. S. H. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP,” in Advances in Neural Information Processing Systems 33, 2020, pp. 9459–9474.

[7] T. Gao et al., “Enabling large language models to generate text with citations,” in Proc. EMNLP, Singapore, 2023, pp. 6465–6488.