SciHallu: A Multi-Granularity Hallucination Detection Dataset for Scientific Writing

Adiba Ibnat Hossain
Northern Illinois University
DeKalb, IL, USA
ahossain4@niu.edu

Sagnik Ray Choudhury
University of North Texas
Denton, TX, USA
sagnik.raychoudhury@unt.edu

Hamed Alhoori
Northern Illinois University
DeKalb, IL, USA
alhoori@niu.edu

Abstract

Large Language Models (LLMs) are increasingly used to support scientific writing, but their tendency to produce hallucinated content threatens academic reliability. Existing benchmarks have addressed hallucination detection in general-domain tasks, such as fact-checking or question answering, but they do not reflect the fine-grained, domain-specific needs of scientific communication. We introduce **SciHallu**, a dataset for identifying hallucinations in academic text at three levels of granularity: token, sentence, and paragraph. To establish a reliable ground truth, we select source passages from research papers published prior to the widespread adoption of LLMs. Our dataset includes both hallucinated and non-hallucinated paragraph instances, constructed through controlled perturbations at varying levels of noise and validated by human annotators. A rationale is paired with each instance, explaining the nature of the modification. SciHallu covers multiple academic fields, such as Computer Science, Health Sciences, and Humanities and Social Sciences. It is built using a model-guided annotation pipeline, followed by expert human validation. We evaluate state-of-the-art LLMs on both binary and fine-grained classification tasks, revealing challenges in detecting subtle hallucinations. SciHallu supports the development of context-aware systems for more trustworthy scientific content generation.

Dataset & Code

1 Introduction

LLMs are reshaping the landscape of communication. They are becoming increasingly popular tools because of their ability to produce fluent, coherent text with little human effort. However, this fluency carries some risks. LLMs are known to produce hallucinations, which refer to content that is factually incorrect, unverifiable, or contextually inconsistent with the original source material (Filippova, 2020).

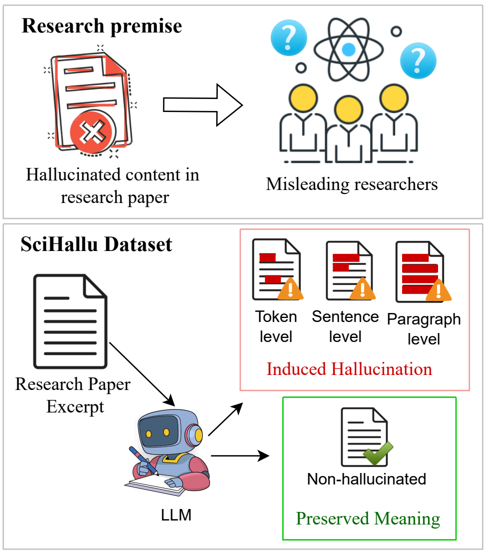

Figure 1: Overview of the SciHallu dataset construction process from scientific papers.

In high-stakes fields like science and health, such inaccuracies can weaken the credibility of scholarly publications, compromise the reproducibility of findings, and distort the cumulative knowledge that future research depends on.

Evidence of widespread LLM use in academic writing is growing. Studies have documented stylistic shifts in scientific publications that correspond with the rise of generative AI tools (Lazebnik and Rosenfeld, 2024; Kobak et al., 2025). This trend is especially prevalent in shorter publications, competitive research areas, and among authors who frequently submit preprints (Liang et al., 2024a). Notably, models like ChatGPT can produce abstracts so convincing that reviewers misclassify them nearly one-third of the time (Gao et al., 2023). This difficulty in detection raises serious concerns, as faulty AI-generated content may reach the scientific literature, earn undeserved credibility, and ultimately affect the direction of future research.

Empirical studies increasingly document specific

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1277–1304
December 20-24, 2025 ©2025 Association for Computational Linguistics

cases in LLM-generated academic text. These include fabricated references, factual distortions, and statements that conflict with established evidence across a range of scientific domains (Buchanan et al., 2025; Ravichander et al., 2025; Oladokun et al., 2025; Walters and Wilder, 2023; Athaluri et al., 2023). Recent studies reveal that these models frequently disseminate misinformation (Alvero, 2023; Li et al., 2024), and produce outputs that deviate from the intended context of scholarly work (Liu et al., 2025). For example, GPT-based models have shown alarmingly low precision when tasked with literature synthesis (as low as 0% with BARD) (Chelli et al., 2024), and even fabricated non-existent facts in medical reports (Alkaissi and McFarlane, 2023). A large-scale empirical analysis further revealed a 17.5% increase in AI-modified content within Computer Science publications, with similar trends observed in Electrical Engineering and Systems Science (Liang et al., 2024a). These trends highlight the critical need for strong hallucination detection systems designed for academic writing, where contextual fidelity and factual accuracy are crucial (Mittelstadt et al., 2023).

Despite increased attention to the problem, existing resources for hallucination detection fall short in two critical ways. First, most datasets use binary labels (e.g., hallucinated vs. non-hallucinated, support vs. refute), which oversimplifies the complex, layered nature of hallucinations that may occur at the level of individual tokens, isolated sentences, or entire paragraphs. Second, most benchmarks are based on general-domain content, such as Wikipedia or news articles, which lack the density, structure, and domain-specific reasoning found in scientific writing. Existing benchmarks such as HaDeS (Liu et al., 2021), HaluEval (Li et al., 2023), and SciFACT (Wadden et al., 2022) are valuable for general tasks like fact-checking and QA, but do not address the fine-grained detection needs specific to scientific discourse.

We specifically designed a dataset for detecting hallucinations in academic writing at various granularities in order to overcome these limitations. Our dataset contains both hallucinated and non-hallucinated instances, with hallucinations categorized into three levels: token-level, sentence-level, and paragraph-level. Each instance is labeled accordingly, resulting in a four-class annotation scheme. Token-level hallucinations introduce words and phrases that don't make sense or are semantically irrelevant; sentence-level hallucinations include false claims or contextually incongruent statements; and paragraph-level hallucinations exhibit broader inconsistencies, like thematic drift or significant factual changes.

To further capture the variability in hallucination severity, we modulate the intensity of perturbations by applying controlled levels of noise across paragraph instances. In addition, human annotators review and validate a subset of the data to ensure consistency with the intended perturbation types and to support high-quality benchmarking.

2 Related Work

Despite their transformative capabilities, LLMs remain prone to hallucinations, generating fluent but factually incorrect, unverifiable, or misleading content due to their probabilistic nature (Hamid, 2024). This challenge has been widely studied in traditional natural language generation (NLG) tasks such as open-ended summarization (Rehman et al., 2024; Shakil et al., 2024; Wan et al., 2024), question-answering (Chen et al., 2023; Snyder et al., 2024), machine translation (Gongas et al.; Guerreiro et al., 2023), and dialogue systems (Dziri et al., 2021; Sun et al., 2023). Previous studies typically rely on general-domain corpora and adopt simplistic binary labels, which may not capture the full spectrum of semantic distortions.

The lack of in-depth domain-specific knowledge is a major limitation of general-domain LLMs trained on web-scale corpora (Raffel et al., 2020), which leaves them vulnerable to generalization bias. When applied to scientific research, this knowledge gap may result in unsubstantiated or misleading outputs (Peters and Chin-Yee, 2025). Without access to validated, domain-specific knowledge sources, LLMs struggle to produce accurate, verifiable responses, posing risks to users and practitioners.

To mitigate hallucinations, several benchmark datasets have been proposed, primarily for evaluating factual consistency. In open-domain QA, datasets like PopQA (Mallen et al., 2023) and TruthfulQA (Lin et al., 2022) evaluate how well an LLM can produce accurate answers to questions on a wide range of subjects. HalluQA (Cheng et al., 2023) and UHGEval (Liang et al., 2024b) assess hallucinations in Chinese LLMs, while FEVEROUS (Aly et al., 2021) expands fact verification to both structured and unstructured data. However, these resources focus on general content and lack the specificity required to evaluate hallucinations in specialized, technical writing.

To improve factual grounding, domain-specific datasets have emerged across fields such as medicine, law, and finance. For example, RadGraph (Jain et al., 2021) supports clinical information extraction from radiology reports, LEDGAR (Tuggener et al., 2020) and LeDQA (Liu et al., 2024) address legal QA tasks, and FinQA (Chen et al., 2021) targets financial reasoning through expert-labeled QA pairs. While these datasets enrich domain knowledge and reduce hallucination rates indirectly, they are not explicitly designed to detect hallucinations or assess multi-level informational reliability in generated text.

A smaller number of datasets focus directly on hallucination detection within domain-specific contexts. PubHealthTab (Akhtar et al., 2022) evaluates tabular claim verification in the public health domain. SciFact (Wadden et al., 2022) provides labeled scientific claims supported or refuted by abstracts. SciHal (Li et al., 2025) collects hallucinated assertions produced by AI research assistants in response to scientific questions. These datasets mark important progress toward domain-sensitive evaluation, but they generally treat hallucination at the sentence or claim level, leaving broader structural inconsistencies underexplored. Furthermore, hallucination detection within academic writing, beyond isolated QA or summarization tasks, remains largely unaddressed.

Another critical limitation of existing work is the lack of fine-grained annotation. Most hallucination detection benchmarks operate at a single level of granularity, typically token-level (Marfurt and Henderson, 2022; Choi et al., 2023; Guo et al., 2024) or sentence-level (Wang et al., 2020; Chen et al., 2024). For example, HADES (Liu et al., 2021) and RAGTruth (Niu et al., 2024) annotate token-level errors, while MHaluBench (Chen et al., 2024) evaluates specific claims within long-form responses. Some work extends to paragraph-level analysis (Zhou et al., 2025), but to the best of our knowledge, no existing dataset systematically captures hallucinations at multiple granular levels within the same task. This is a significant gap, as hallucinations in scientific writing often occur at multiple linguistic and conceptual layers, from individual terminology errors to high-level thematic drift.

Moreover, the structured rhetoric, use of technical jargon, and high factual complexity of scientific writing present unique difficulties. Hallucinations of this genre are not only harder to detect but may also have more serious consequences. Thus, there is a clear need for a benchmark tailored to this domain, one that supports fine-grained, hierarchical hallucination detection in research-oriented text.

3 Methodology

To build a robust dataset for hallucination detection in scientific writing, we systematically introduced hallucinations into research text by perturbing content at the token, sentence, and paragraph levels. These perturbations were designed to simulate varying degrees of content distortion, ranging from subtle distortions to severe semantic deviations, depending on the assigned hallucination type and noise intensity. In addition to hallucinated **variants** (modified versions of the original paragraph), we generated non-hallucinated counterparts by applying minimal edits that preserved the original meaning and structure. To ensure the quality and diversity of the variants, we generated multiple versions of each paragraph. We evaluated them using a set of scoring metrics and selected the most representative variant for inclusion. This process was followed by model-assisted annotation and expert human validation. Figure 2 illustrates the overall pipeline, elaborated on in the following subsections.

3.1 Data Collection

To construct the dataset on a reliable foundation, we selected research papers published before the widespread adoption of LLMs and focused on papers released between 2018 and 2021. We chose this pre-LLM time frame to minimize the possibility of LLM-generated content in the original texts. The dataset includes papers from three major domains: computer science, humanities and social sciences, and health sciences, where each **paragraph** from these papers was treated as the primary unit for variant generation and analysis. Table 6 summarizes the distribution of collected data across domains and publication venues, including the number of source papers and paragraph instances.

We extracted 66,552 paragraph instances from 2,377 research papers. Each **instance** includes a paragraph along with relevant metadata such as domain, venue, title, abstract, and section. To enhance contextual understanding during downstream processing, we also captured the preceding and following paragraphs for each instance.

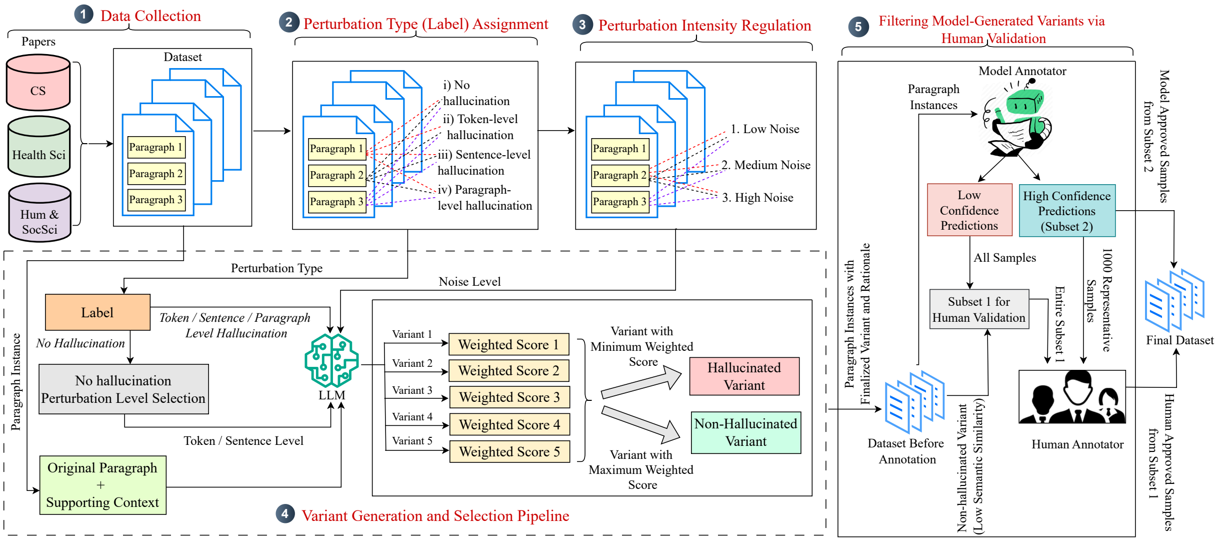

Figure 2: SciHallu dataset construction pipeline

3.2 Perturbation Type Assignment

To guide variant generation, we first randomly assigned each instance a perturbation type (label). To avoid overrepresentation of any single venue within a given data partition, we began by randomly shuffling the dataset. We then split the data into four equal subsets to maintain class balance. Each subset was assigned one of the following perturbation types: (1) *No hallucination*, (2) *Token-level hallucination*, (3) *Sentence-level hallucination*, or (4) *Paragraph-level hallucination*, resulting in 16,638 instances per category. After labeling, we re-shuffled the data to ensure a randomized distribution of perturbation types throughout the dataset.

3.3 Perturbation Intensity Regulation

To control the severity of hallucinations introduced during perturbation, we applied a noise-level mechanism that assigns each paragraph instance a random noise level from 1 to 3. These levels represent the degree of semantic distortion, as follows:

- **Level 1 - Low Noise:** Subtle substitutions with minimal impact on the core meaning of the paragraph.
- **Level 2 - Medium Noise:** Edits that introduce noticeable hallucinations, causing partial misinformation or mild semantic drift that could mislead an inattentive reader.
- **Level 3 - High Noise:** Substantial modifications that significantly distort the meaning of the paragraph, resulting in high semantic deviation and increased risk of misinterpretation.

3.4 Variant Generation and Selection Pipeline

3.4.1 Controlled Variant Generation

As part of the SciHallu dataset construction, we aimed to produce perturbed paragraphs that represent varying degrees of hallucination at different granularities. Each generation was guided by a predetermined perturbation label that specified whether the output should introduce hallucination at the token, sentence, or paragraph level or stay true to the source.

For every original paragraph, we produced multiple candidate variants to guarantee the inclusion of high-quality representative examples. While hallucinated variants were produced by introducing controlled perturbations, non-hallucinated variants were generated through minor and semantically neutral edits designed to preserve the original meaning and structure of the paragraph. For these non-hallucinated cases, we randomly selected either the token or sentence level as the basis for subtle edits, with care taken to avoid semantic distortion. Including these minimally altered non-hallucinated variants keeps the variant generation process consistent across both classes and helps models learn to distinguish between benign textual variations and true hallucinations.

We produced five candidate variants per paragraph using GPT-4o-mini (OpenAI, 2024), a model shown to be effective for synthetic dataset creation (Hastings et al., 2024; Kim et al., 2024). This model has also been adopted in recent benchmarks for similar generation tasks (Workum et al., 2025; Krechetova and Kochedykov, 2025). Each generation was prompted with detailed instructions that included the original paragraph, relevant context, the assigned perturbation type and its definition, as well as the designated noise level. For non-hallucinated variants, the noise level was set to 0 to ensure semantic preservation.

The context provided to the model varied depending on the targeted hallucination granularity. For non-hallucinated (token-level edits) (Figure 12) and token-level hallucination generation (Figure 9), we supplied the title and abstract to provide localized context, which is typically sufficient for guiding small-scale edits. For non-hallucinated (sentence-level edits) (Figure 13), sentence- and paragraph-level hallucinations (Figures 10, 11), we included the adjacent paragraphs to offer a broader context window. This additional context helps the model better understand the boundaries within which it can introduce meaningful perturbations. Sentence-level hallucinations often conflict with the local context of the paragraph, while paragraph-level hallucinations tend to deviate from the global discourse or thematic flow. By giving the model an extended context, we encourage it to generate hallucinations that are more coherent and context-aware and align with the desired perturbation level.

Sentence-level hallucinations usually involve subtle changes in content that are only noticeable when compared to the neighboring text. On the other hand, hallucinations that occur at the paragraph level cause more significant factual or thematic deviations throughout the discourse. We did not constrain the number of edits per variant, allowing for natural variation in the extent and form of modification. Based on empirical tuning, we set the model's temperature to 0.7 and top-p value to 0.9. Temperature regulates how randomly tokens are chosen; higher values promote more diverse and creative results. To minimize the possibility of incoherent output while maintaining diversity, top-p (nucleus sampling) restricts the model to sampling from the smallest set of tokens whose cumulative probability exceeds p (Holtzman et al., 2019).

Each generated variant was accompanied by a *rationale* produced by the model. These rationales explain the nature of the modification and how it aligns with the assigned perturbation type. They improve interpretability, enhance transparency, and document the transformation process. We also computed evaluation metrics for each variant to facilitate the selection process described in Subsubsection 3.4.2. Each final paragraph instance selected for inclusion in SciHallu contains the fields described in Table 1.

Table 1: Structure of dataset instances and associated metadata fields.

Field	Description
Domain	Subject area of the paper.
Venue	Conference or journal where the paper was published.
Title	Title of the research paper.
Abstract	Summary of the paper content.
Section	Section title from which the paragraph was extracted.
Previous paragraph	Paragraph preceding the original paragraph.
Original paragraph	Central paragraph used as the basis for variant generation.
Next paragraph	Paragraph following the original paragraph.
Modified paragraph	Perturbed version of the original paragraph.
Rationale	Model-generated explanation describing the modification and its type.
Noise	Assigned noise level (0: none to 3: high).
Perplexity	Fluency score of the modified paragraph.
Similarity (orig.–mod.)	Semantic similarity between original and modified paragraphs.
Similarity (abs.–mod.)	Semantic similarity between abstract and modified paragraphs.
Label	Hallucination type: none, token-, sentence-, or paragraph-level.

3.4.2 Metric-Driven Variant Selection

To select the most suitable candidate for each instance, we ranked the generated variants based on a composite score derived from three evaluation metrics:

(1) Perplexity: Lower perplexity indicates higher fluency and naturalness, ensuring that both hallucinated and non-hallucinated variants sound coherent. We prioritized variants with lower perplexity in both categories to favor fluent generation.

(2) Semantic similarity between the original and modified paragraph: This metric helps capture the extent of content distortion. For hallucinated variants, we expect lower similarity, reflecting intentional deviations from the original. For non-hallucinated variants, higher similarity indicates semantic preservation.

(3) Semantic similarity between the abstract and modified paragraph: For semantic alignment, this serves as an auxiliary signal. While non-hallucinated variants stay semantically close to the abstract, hallucinated ones drift further away.

Table 2 summarizes the preferred directional patterns of each evaluation metric when selecting hallucinated and non-hallucinated variants.

Table 2: Preferred directions of evaluation metrics.

Metric	Hallucinated	Non-hallucinated
Perplexity	↓	↓
Similarity (Original–Modified)	↓	↑
Similarity (Abstract–Modified)	↓	↑

For consistent scoring, we applied min-max normalization to perplexity values using the global minimum and maximum observed across the dataset, scaling them to the range [0,1]. We then computed a weighted score by combining the three metrics with equal weights (0.33 each). Equal weighting ensures general applicability across topics and domains, preventing any single metric from dominating the evaluation.

For **hallucinated variants**, optimal cases correspond to lower values across all three metrics. We therefore selected the variant with the lowest weighted score (Table 7, 8, 9), computed according to equation 1:

$$W_{\text{hallu}} = \frac{1}{3}\left(\text{NP} + S_{\text{orig}} + S_{\text{abs}}\right) \tag{1}$$

where W_{hallu} = Weighted score for hallucinated variant, NP = Normalized perplexity, S_{orig}= Semantic similarity with the original paragraph, and S_{abs}= Semantic similarity with the abstract.

For **non-hallucinated variants**, the preferred metric configuration involves low perplexity and high semantic similarity to both the original paragraph and the abstract. To align the direction of all metrics (so that higher values indicate better candidates), we transformed the perplexity score by computing its complement: (1 - NP). This transformation allows all metrics to contribute positively to the overall score. We then selected the non-hallucinated variant with the highest weighted score (Table 10), calculated as:

$$W_{\text{nonhallu}} = \frac{1}{3}\left((1 - \text{NP}) + S_{\text{orig}} + S_{\text{abs}}\right) \tag{2}$$

where W_{nonhallu} is the weighted score for non-hallucinated variants.

3.5 Filtering Model-Generated Variants via Human Validation

We used a two-stage validation process that combined model-assisted filtering and expert human review to verify the reliability of the generated data. The main objective was to confirm whether each modified paragraph aligned with its assigned perturbation type and whether the accompanying rationale accurately described the modification.

Given the large dataset size, we first applied uncertainty-based sampling (Raj and Bach, 2022) to prioritize which samples required human evaluation. We used the DeepSeek-7B model (DeepSeek-AI et al., 2024) to perform binary classification on each instance, determining whether the modified paragraph and its rationale were satisfactory. The model outputs a “yes” or “no” decision, along with a confidence score derived from its logit values (Figure 6). To identify cases that required closer inspection, we created two evaluation subsets:

- **Subset 1 (Low-Confidence and Outlier Variants):** This group included (a) all variants (hallucinated and non-hallucinated) where the model’s classification confidence was low, and (b) non-hallucinated variants with unexpectedly low semantic similarity to the original paragraph (similarity score < 0.9, flagged as outliers using the lower whisker method; Figure 5). Using the formula in Equation 3, we defined a “low confidence” threshold of 87.85%.

$$\text{Threshold} = \text{Mean} - 2 \times \text{Std} \tag{3}$$

  We filtered 3,407 paragraph variants into Subset 1 for expert validation.

- **Subset 2 (High-Confidence Variants):** The remaining paragraph variants with high model confidence were placed in this group. To assess the reliability of the model’s judgments, we randomly sampled 1,000 variants from Subset 2, maintaining a balanced 50:50 split between model-labeled “yes” and “no” categories.

Three expert annotators (computer scientists with relevant domain knowledge) independently reviewed each variant, evaluating both the modified paragraph and its rationale. For Subset 1,129 of the 3,407 variants were flagged as inadequate (majority "no" vote) and removed. The remaining 3,278 were retained as validated paragraph instances. Inter-annotator agreement, measured using Fleiss' Kappa, was strong ($\kappa = 0.895$). For Subset 2, annotators agreed with the model in 96.6% of the reviewed representative variants, with a Fleiss' Kappa of 0.913, indicating excellent agreement. This result confirmed the credibility of the model's high-confidence judgments.

To finalize the dataset, we excluded all high-confidence variants flagged as "no" by the model. For non-hallucinated entries appearing in both subsets (due to low similarity), we retained the human-approved instances from Subset 1 and removed duplicates from Subset 2 to maintain consistency. After filtering and consolidation, the final dataset consisted of 62,596 paragraph instances (Table 3).

4 Dataset Analysis and Insights

The final SciHallu dataset consists of paragraph instances spanning three academic domains. Table 3 provides a breakdown of instance counts per hallucination type and domain in the resulting dataset.

Table 3: Instance statistics in SciHallu across domains

Label	Domain		
	CS	Hum & SocSci	Health Sci
No Hallucination	12,466	2,833	1,657
Token-level Hallucination	11,590	2,597	1,695
Sentence-level Hallucination	10,700	2,633	1,586
Paragraph-level Hallucination	10,762	2,537	1,540

We observed the semantic similarity and perplexity metrics of the generated variants to get a better idea of their quality and characteristics. Figure 3 illustrates how these two indicators vary across hallucination types. Non-hallucinated instances consistently show low perplexity and high semantic similarity to the original paragraphs. In contrast, paragraph-level hallucinations display broader variation and semantic drift. Furthermore, some token-level hallucinated variants show increased perplexity, indicating decreased fluency.

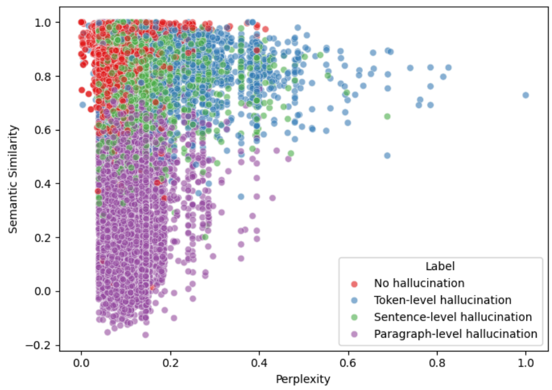

Figure 3: Similarity vs. perplexity across labels

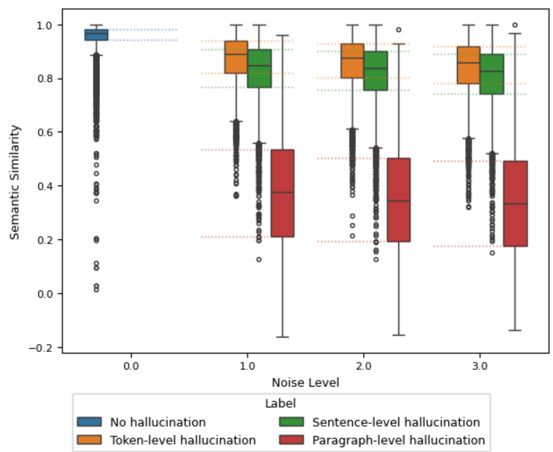

Figure 4: Semantic similarity vs. noise level for labels

We also examined how the semantic similarity between original and modified paragraphs changes with different noise levels and hallucination types. Figure 4 shows that variants with paragraph-level hallucination diverge the most from their originals, whereas most non-hallucinated instances retain high similarity. On the other hand, hallucinated instances with lower noise intensity exhibit increased semantic similarity. This trend highlights how semantic deviation scales with the intensity and granularity of perturbation.

4.1 Evaluating LLM Performance on Hallucination Detection

We evaluated the performance of six LLMs, Mistral-7B (Jiang et al., 2023), LLaMA2-7B (Touvron et al., 2023), LLaMA3-8B (Grattafiori et al., 2024), DeepSeek-Qwen-8B (DeepSeek-AI, 2025), Qwen3-14B (Yang et al., 2025), and Phi-4-14B (Abdin et al., 2024) on our dataset to explore their

Table 4: Model accuracy (%) of binary hallucination classification task.

Model Name	Label-wise Accuracy				Overall Accuracy
	No hallucination	Token-level hallucination	Sentence-level hallucination	Paragraph-level hallucination	
mistral-7b	**96.97**	26.32	35.53	49.11	53.05
llama2-7b	74.56	32.50	32.91	30.19	43.44
llama3-8b	69.53	40.16	41.87	41.39	48.81
deepseekqwen-8b	7.16	**95.06**	**95.62**	**91.10**	70.44
qwen3-14b	51.72	89.59	92.29	82.46	78.28
phi4-14b	74.25	81.81	86.18	87.42	**82.13**

Table 5: Model accuracy (%) of fine-grained hallucination classification task.

Model Name	Label-wise Accuracy				Overall Accuracy
	No hallucination	Token-level hallucination	Sentence-level hallucination	Paragraph-level hallucination	
mistral-7b	49.61	**2.34**	60.06	38.10	37.38
llama2-7b	17.80	0.89	3.83	**82.09**	25.42
llama3-8b	55.24	0.28	25.72	60.95	35.61
deepseekqwen-8b	25.25	0.14	81.10	0.02	26.21
qwen3-14b	31.24	0.48	**89.78**	53.46	**42.66**
phi4-14b	**89.05**	0.30	6.43	19.81	30.43

capacity for detecting hallucinations in scientific texts. We explored the following two questions:

4.1.1 Can LLMs detect hallucination in scientific literature?

To explore whether LLMs can identify the presence of hallucination, we prompted each model to perform binary classification using the format shown in Figure 7. Each paragraph variant was labeled as either "YES" (if any hallucination was present, regardless of granularity) or "NO" (if no hallucination was present). All hallucinated variants (token-level, sentence-level, or paragraph-level) were treated as belonging to a unified hallucinated class. Furthermore, during evaluation, we disaggregate the hallucinated class to assess model sensitivity to different types of hallucinations and observe the label-wise accuracy. As summarized in Table 4, most models struggle with this task. While Mistral-7B performs best on non-hallucinated instances, DeepSeek-Qwen-8B shows the strongest performance across the hallucinated categories.

4.1.2 Can LLMs detect hallucination at a granular level in scientific literature?

We then evaluated whether LLMs can distinguish between different hallucination granularities. Using a multi-class zero-shot setting (prompt shown in Figure 8), we asked each model to classify paragraph variants into one of four categories: non-hallucinated, token-level, sentence-level, or paragraph-level hallucination. As shown in Table 5, model performance drops significantly in this more fine-grained setting. Token-level hallucinations were the hardest for all models to detect, likely due to their subtle nature and lack of clear contextual disruption. In contrast, hallucinations at the sentence and paragraph levels, where content deviates more explicitly from the original, were comparatively easier for the models to identify. The results imply that, while existing LLMs are somewhat capable of identifying overt hallucinations, fine-grained detection is still difficult to achieve.

5 Conclusion

We present SciHallu, the first dataset created to aid the detection and mitigation of hallucinations in scientific writing. SciHallu, built on controlled perturbations of scholarly content, allows for fine-grained analysis across various hallucination levels and academic domains. Our evaluation of state-of-the-art LLMs shows that the models struggle in correctly identifying hallucinations in scientific texts, particularly when they are subtle or occur at lower granularities. These results point to a critical gap in the current LLM capabilities and emphasize the necessity of specialized resources to deal with the particular difficulties associated with scientific hallucination. SciHallu offers a domain-grounded, scalable, and interpretable benchmark paving the way for the creation of more trustworthy, contextually aware models that can preserve semantic integrity in intricate scientific discourse.

Limitations

SciHallu provides a scalable and fine-grained benchmark for detecting hallucinations in scientific literature, but there are still significant limitations that create potential for future work. First, the dataset is constructed from synthetically induced hallucinations rather than real-world outputs from LLM-generated scientific articles. The complete intricacy and ambiguity of naturally occurring hallucinations are absent from synthetic generation, despite the fact that it allows controlled perturbations, validated labeling, and rationale-grounded interpretability. Incorporating genuine or semi-synthetic hallucinations could bridge this gap, but it raises issues such as verification subjectivity and ethical concerns about mislabeling peer-reviewed information. Our current design mitigates these concerns while enabling rigorous benchmarking, though we consider hybrid data integration an important future direction.

Second, SciHallu currently includes content from only three academic domains: Computer Science, Health Sciences, and Humanities & Social Sciences. This limited coverage could have an impact on the generalizability of trained detection models to other fields like Natural Sciences, Engineering, or Economics, where factual density, discourse patterns, and citation practices can differ substantially. Increasing robustness and domain transferability will require expanding to underrepresented fields.

Third, while similarity scores and perplexity serve as useful signals for filtering and ranking candidate variants, they are not definitive indicators of hallucination. High similarity does not rule out subtle factual distortions, and low similarity may capture valid paraphrasing. To address this, our pipeline incorporates human-validated rationales and perturbed texts to ensure that hallucinations are defined not by form but by their lack of grounding in the original source.

Ethical Considerations

This work involves the generation and detection of hallucinations in scientific writing using large language models (LLMs). To avoid ethical risks associated with mislabeling or misinterpreting existing scientific publications, we did not attempt to identify hallucinations in real-world research papers. Instead, all hallucinated content in SciHallu was synthetically created by applying controlled perturbations to original source paragraphs, allowing us to maintain full transparency and label accuracy. Human annotations were conducted only on these synthetic or perturbed texts, with annotators instructed to assess alignment with the original paragraph. No sensitive or personally identifiable information was included in the dataset construction.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2022. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1–16, Seattle, United States. Association for Computational Linguistics.

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

Ruben Alvero. 2023. Chatgpt: rumors of human providers' demise have been greatly exaggerated. *Fertility and Sterility*, 119(6):930–931.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Sai Anirudh Athaluri, Sandeep Varma Manthena, VSR Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Rama Tulasi Siri Duddumpudi. 2023. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references. *Cureus*, 15(4).

Joy Buchanan, Olga Shapoval, and Stephen Hill. 2025. Llm hallucination of citations in economics persists with web-enabled models. *Available at SSRN 5318314*.

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. 2024. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: comparative analysis. *Journal of medical Internet research*, 26:e53164.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252, Bangkok, Thailand. Association for Computational Linguistics.

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.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and 1 others. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. KCTS: Knowledge-constrained tree search decoding with token-level hallucination detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14035–14053, Singapore. Association for Computational Linguistics.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, and 69 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.

Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2023. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *NPJ digital medicine*, 6(1):75.

Laura Gongas, Kenza Benkirane, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, and Eduardo Sánchez. Machine translation hallucination detection for low and high resource languages using large language models. In *Eighth Widening NLP Workshop (WiNLP 2024) Phase II*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Luanxu Guo, Yuanyuan Fang, Feng Chen, Pengcheng Liu, and Song Xu. 2024. Large language models with adaptive token fusion: A novel approach to reducing hallucinations and improving inference efficiency. *Authorea*.

Oussama H Hamid. 2024. Beyond probabilities: Unveiling the delicate dance of large language models (llms) and ai-hallucination. In *2024 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pages 85–90. IEEE.

John D Hastings, Sherri Weitl-Harms, Joseph Doty, Zachary J Myers, and Warren Thompson. 2024. Utilizing large language models to synthesize product desirability datasets. In *2024 IEEE International Conference on Big Data (BigData)*, pages 5352–5360. IEEE.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Nguyn Dng, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2024. Evaluating language models as synthetic data generators.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2025. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813.

Varvara Krechetova and Denis Kochedykov. 2025. Geobenchx: Benchmarking llms for multistep geospatial tasks. *arXiv preprint arXiv:2503.18129*.

Teddy Lazebnik and Ariel Rosenfeld. 2024. Detecting llm-assisted writing in scientific communication: Are we there yet? *Science*, 9(3):1–10.

Dan Li, Bogdan Palfi, Colin Zhang, Jaiganesh Subramanian, Adrian Raudaschl, Yoshiko Kakita, Anita De Waard, Zubair Afzal, and Georgios Tsatsaronis. 2025. Overview of the SciHal25 shared task on hallucination detection for scientific content. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 307–315, Vienna, Austria. Association for Computational Linguistics.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

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.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, and 1 others. 2024a. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*.

Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Yezhaohui Wang, Dawei He, Cheng Peng, Zhonghao Wang, and Haiying Deng. 2024b. UHGEval: Benchmarking the hallucination of Chinese large language models via unconstrained generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5266–5293, Bangkok, Thailand. 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.

Bulou Liu, Zhenhao Zhu, Qingyao Ai, Yiqun Liu, and Yueyue Wu. 2024. Ledqa: A chinese legal case document-based question answering dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5385–5389.

Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. 2025. Towards long context hallucination detection.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*.

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.

Andreas Marfurt and James Henderson. 2022. Unsupervised token-level hallucination detection from summary generation by-products. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 248–261, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. To protect science, we must use llms as zero-shot translators. *Nature human behaviour*, 7(11):1830–1832.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Bolaji David Oladokun, Rexwhite Tega Enakrire, Adefila Kolawole Emmanuel, Yusuf Ayodeji Ajani, and

Adebowale Jeremy Adetayo. 2025. Hallucitation in scientific writing: Exploring evidence from chatgpt versions 3.5 and 4o in responses to selected questions in librarianship. *Journal of Web Librarianship*, 19(1):62–92.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. OpenAI blog.

Uwe Peters and Benjamin Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776.

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.

Anant Raj and Francis Bach. 2022. Convergence of uncertainty sampling for active learning. In *International conference on machine learning*, pages 18310–18331. PMLR.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. 2025. Halogen: Fantastic llm hallucinations and where to find them. *arXiv preprint arXiv:2501.08292*.

Tohida Rehman, Ronit Mandal, Abhishek Agarwal, and Debarshi Kumar Sanyal. 2024. Hallucination reduction in long input text summarization. In *International Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023)*, pages 307–316. CRC Press.

Hassan Shakil, Zeydy Ortiz, Grant C Forbes, and Jugal Kalita. 2024. Utilizing gpt to enhance text summarization: A strategy to minimize hallucinations. *Procedia Computer Science*, 244:238–247.

Ben Snyder, Marius Moisescu, and Muhammad Bilal Zafar. 2024. On early detection of hallucinations in factual question answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2721–2732.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13618–13626.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

WH Walters and EI Wilder. 2023. Fabrication and errors in the bibliographic citations generated by chatgpt. sci rep 13, 14045.

David Wan, Koustuv Sinha, Srini Iyer, Asli Celikyilmaz, Mohit Bansal, and Ramakanth Pasunuru. 2024. Acueval: Fine-grained hallucination evaluation and correction for abstractive summarization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10036–10056.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Jessica D Workum, Bas WS Volkers, Davy van de Sande, Sumesh Arora, Marco Goeijenbier, Diederik Gommers, and Michel E van Genderen. 2025. Comparative evaluation and performance of large language models on expert level critical care questions: a benchmark study. *Critical Care*, 29(1):72.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. 2025. Hademif: Hallucination detection and mitigation in large language models. In *The Thirteenth International Conference on Learning Representations*.

A Appendix

A.1 Data Sources

An overview of the data used to construct the Sci-Hallu dataset is presented in Table 6. It includes the distribution of paragraph instances across three academic domains: Computer Science, Health Sciences, and Humanities & Social Sciences, as well as the associated publication venues. These source papers form the foundation for generating both hallucinated and non-hallucinated variants used throughout the dataset.

Table 6: Overview of data sources by domain and publication venue.

Domain	Venue	Papers	Paragraph Instances
CS	ACL 2020	448	14,721
CS	CVPR 2020	473	16,394
CS	NuerIPS 2020	498	1,7907
Hum & SocSci	ICOLLITE 2018	88	1,889
Hum & SocSci	ICOLLITE 2019	72	1,666
Hum & SocSci	ICOLLITE 2020	130	3,043
Hum & SocSci	ICOLLITE 2021	116	3,015
Hum & SocSci	ICLCCS 2020	70	1,185
Health Sci	ICCvD 2021	52	811
Health Sci	AHMS 2020	65	1,012
Health Sci	ICoSIHSN 2020	116	1,739
Health Sci	ICHD 2020	80	936
Health Sci	ICCSCP 2021	50	717
Health Sci	ICoSHEET 2019	119	1,517
Total		**2,377**	**66,552**

A.2 Experiment Setup for Variant Generation

To construct the dataset, we generated multiple paragraph-level variants using prompt templates tailored to different hallucination types. Templates for token-, sentence-, and paragraph-level hallucinations are shown in Figures 9, 10, and 11, respectively. For non-hallucinated cases, we applied minimal edits at either the token level (Figure 12) or the sentence level (Figure 13) to simulate minor, semantically neutral modifications. For each original paragraph, five candidate variants were generated. From these, we selected the most representative one to include as the final dataset instance. After selection, we inspected the resulting instances to assess their quality and diversity, where we noticed that some non-hallucinated instances showed unexpectedly low semantic similarity to their source paragraphs. As illustrated in Figure 5, these outliers were flagged for further human validation.

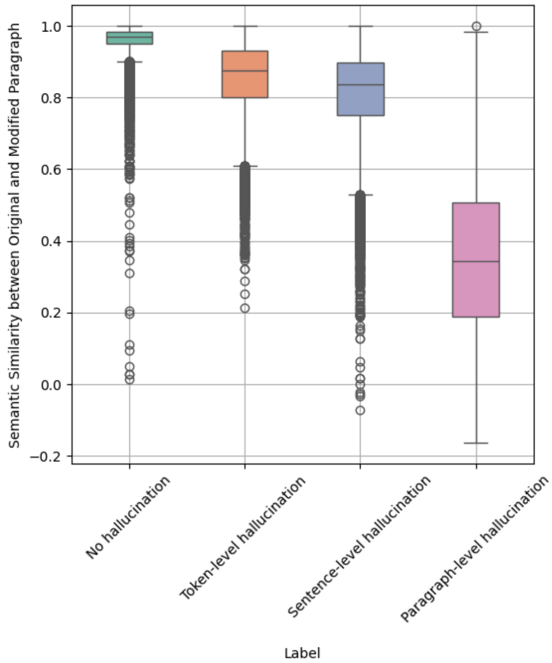

Figure 5: Semantic consistency across hallucination categories.

A.3 Illustrative Samples for All Classification Categories

Table 7, 8, 9, and 10, respectively, present five generated variants along with the selected sample for token, sentence, paragraph-level hallucinations, and no hallucination categories.

A.4 Model-assisted Annotation Setup

We instructed DeepSeek-7b with the prompt mentioned in Figure 6 for primary judgment.

A.5 Human Annotation

Human annotators were provided with the following instructions to judge the low-confidence variants (both hallucinated and non-hallucinated) and low-semantic-similarity non-hallucinated variants.

Instructions: You will be presented with a collection of data related to scholarly articles. The data contains hallucinated and non-hallucinated variants obtained by perturbing scientific texts. You need to determine whether the modifications align with the provided label and whether the explanation adequately explains the modification. Please carefully review all of the information before selecting the YES or NO response to each of

You are a judge who can analyze text from research papers with the help of good reasoning. I will give you
Title: {title}
Abstract: {abstract}
Original Paragraph: {paragraph}
Modified Paragraph: {modified_paragraph}
Explanation: {explanation}} where it is mentioned how the original paragraph has been modified and why the modified paragraph introduces {label}. {definition _of_label}
You need to ONLY answer YES or No based on the following questions.

1. Does the modified paragraph introduce {label}?
2. Does the explanation clearly mention what modification are made in the original paragraph to introduce hallucination?

If both of the answers are YES for both of the questions, respond YES. Otherwise say NO. Do not add any explanation.

Figure 6: Prompt used to instruct the model to judge the modified paragraphs and their corresponding rationales.

the evaluation questions. If all the answers to the questions are YES, put it as your final answer; otherwise, NO.

Provided Information (for each variant): Title, abstract, original paragraph, modified paragraph, rationale of modification, label, and definition of label.

Evaluation Questions

- Does the modified paragraph introduce the intended perturbation? [YES/NO]
- Does the rationale clearly describe the modifications? [YES/NO]
- Does the modified **non-hallucinated** variant contain any sort of hallucination? [YES/NO]
- Does the semantic deviation of the non-hallucinated variant from the original paragraph contradict the intended modification objective? [YES/NO]
- Should this instance be removed from the dataset? [YES/NO]

A.6 Model Evaluation on SciHallu

We evaluated six LLMs on both binary and fine-grained classification tasks, using the prompts shown in Figure 7 and Figure 8, respectively.

You are given a research paper excerpt and determine if the paragraph contains hallucinations (factual inaccuracies / unsupported claims/nonsensical terms):

Title: {title}
Abstract: {abstract}
Context: {previous_paragraph} [PARAGRAPH TO CHECK] {next_paragraph}
Paragraph: {modified_paragraph}

Does this paragraph contain hallucinations? Start the answer strictly with "YES" or "NO".

Figure 7: Prompt to detect hallucination at the binary level.

Analyze this research paper excerpt and classify the paragraph into one of these categories based on hallucination level:
1. NO_HALLUCINATION - No factual or contexual inaccuracies.
2. TOKEN_LEVEL - Nonsensical, incorrect words or phrases.
3. SENTENCE_LEVEL - One or more sentences contain unsupported claims or factual inaccuracies or contextual inaccuracies.
4. PARAGRAPH_LEVEL - The entire paragraph has semantically drifted from the abstract and the context.
Title: {title}
Abstract: {abstract}
Context: {previous_paragraph}
[PARAGRAPH TO CHECK] {next_paragraph}
Paragraph: {modified_paragraph}

Your answer must be EXACTLY one of:
NO_HALLUCINATION, TOKEN_LEVEL, SENTENCE_LEVEL, PARAGRAPH_LEVEL.
Answer:

Figure 8: Prompt to detect hallucination at the fine-grained level.

Token-level Hallucination Generation

[System Input]:
You are given a title, abstract, section and a paragraph of the section from a research paper. Your task is to introduce **token-level hallucination** to the paragraph by replacing a few individual words with unrelated words (unrelated but not absurd words), while keeping the overall structure and most of the sentences intact.

While replacing words, make sure they are not related to the title, abstract and section. Keep the words within the {**given_domain**} domain, but make them **incorrect, misleading, or out-of-context** within that domain. This helps maintain realism while still introducing hallucination.

Do not add or remove entire phrases or sentences, just substitute some existing words with incorrect or out-of-context ones. The substitution of word may include fake named entity, wrong number, fake acronym and many more. Do not substitute stop words. Introduce level {**noise_level**} hallucination in the given paragraph. The hallucination level is defined below:

1. Low Noise (Minor Distortion):

Substitutions are minimal and have little to no impact on the core semantics of the paragraph. The text remains coherent and truthful, with only slight inaccuracies or out-of-place terms.

Barely noticeable, low risk of misunderstanding.

2. Medium Noise (Moderate Distortion):

The paragraph contains noticeable but not overwhelming hallucinations. Key terms may be replaced or altered, introducing partial misinformation or semantic drift.

Misleading without careful reading, moderate impact on factual integrity.

3. High Noise (Severe Distortion):

Substantial hallucination significantly alters or corrupts the meaning of the paragraph. The substitutions make the text misleading, nonsensical, or factually incorrect.

High risk of misunderstanding, major semantic degradation.

Provide the perplexity of the modified paragraph too.

Provide an explanation of why the modified paragraph has token-level hallucination and also mention all the words in the explanation that you modified from the original paragraph. Follow a common structure for explanation in all the responses. Maintain the following template in the output.

Modified_Paragraph:

Perplexity:

Explanation:

[User Input]:
Title: {**title**}, Abstract: {**abstract**}, Section: {**section**}, Paragraph: {**paragraph**}

Figure 9: Prompt template for generating token-level hallucinated variants.

Sentence-level Hallucination Generation

[System Input]:

You are given a title, abstract, section, a current paragraph of the section and its adjacent two paragraphs (previous paragraph and next paragraph) from a research paper.

Your task is to introduce **sentence-level hallucination** to the current paragraph by replacing a few sentences with unrelated sentences (unrelated but not changing the theme of the overall paragraph), while keeping the overall structure of the paragraph intact.

The sentence-level hallucination can affect the logic or local context of the paragraph without changing the theme of the paragraph. The modified paragraph should not deviate from the global context. For global context refer to the title, abstract and adjacent paragraphs (if adjacent paragraphs exist).

While replacing sentences, make sure they are not related to the title, abstract and section. Keep the sentences within the {**given_domain**} domain, but make them incorrect or misleading within that domain. This helps maintain realism while still introducing hallucination.

You MUST NOT add any new sentences. The number of sentences in the modified paragraph must be EXACTLY the same as in the original paragraph. Every replaced sentence must correspond 1-to-1 with an original sentence. The substitution of sentences should introduce logical inconsistency or context inconsistency in the paragraph. The substitution should influence the paragraph's local coherence but should not disrupt the global coherence.

Introduce level {**noise_level**} hallucination in the given paragraph. The hallucination level is defined below:

1. Low Noise (Minor Distortion):

Substitutions are minimal and have little to no impact on the core semantics of the paragraph. The text remains coherent and truthful, with only slight inaccuracies or out-of-context sentences.

Barely noticeable, low risk of misunderstanding.

2. Medium Noise (Moderate Distortion):

The paragraph contains noticeable but not overwhelming hallucinations. Important sentences may be replaced or altered, introducing partial misinformation or semantic drift.

Misleading without careful reading, moderate impact on factual integrity.

3. High Noise (Severe Distortion):

Substantial hallucination significantly alters or corrupts the meaning of the paragraph. The substitutions make the text misleading, nonsensical, or factually incorrect.

High risk of misunderstanding, major semantic degradation.

Provide the perplexity of the modified paragraph too. Provide an explanation of why the modified paragraph has sentence-level hallucination and also mention all the sentences in the explanation that you modified from the original paragraph. Follow a common structure for explanation in all the responses. Maintain the following template in the output.

Modified_Paragraph:

Perplexity:

Explanation:

[User Input]:

Title: {**title**}, Abstract: {**abstract**}, Section: {**section**}, Current Paragraph: {**paragraph**}, Previous Paragraph: {**previous_paragraph**}, Next Paragraph: {**next_paragraph**}

Figure 10: Prompt template for generating sentence-level hallucinated variants.

Paragraph-level Hallucination Generation

[System Input]:
You are given a title, abstract, section, a current paragraph of the section and its adjacent two paragraphs (previous paragraph and next paragraph) from a research paper.

Your task is to introduce **paragraph-level hallucination** to the current paragraph by modifying the entire paragraph that the paragraph's main topic, intention, or argument no longer fits the rest of the paper and the paragraph deviates from the global context.

For global context refer to the title, abstract and adjacent paragraphs (if adjacent paragraphs exist).

If the current paragraph does not have adjacent paragraphs introduce paragraph-level hallucination in such a way that the paragraph does not fit into that particular section, its theme drifts from the title and abstract. The modified paragraph focuses on a new theme.

Keep the total number of sentences in the paragraph same. Paragraph-level hallucination involves changing the main context, focus, key facts, or underlying claims of the paragraph, while maintaining sentence structure and flow.

The hallucinated paragraph must appear plausible and remain within the {**given_domain**} domain, but its overall meaning should become misleading, incorrect, or inconsistent with the title, abstract, adjacent paragraphs and section.

Avoid introducing nonsense or absurd claims. Do not insert or remove sentences—just modify enough of the paragraph's content to create a meaningful semantic deviation. Introduce level {**noise_level**} hallucination as defined below:

1. Low Noise (Minor Distortion):

Substitutions are minimal and have little to no impact on the core semantics of the paragraph. The text remains coherent and truthful, with only slight inaccuracies or out-of-context sentences.

Barely noticeable, low risk of misunderstanding.

2. Medium Noise (Moderate Distortion):

The paragraph contains noticeable but not overwhelming hallucinations. Important sentences may be replaced or altered, introducing partial misinformation or semantic drift.

Misleading without careful reading, moderate impact on factual integrity.

3. High Noise (Severe Distortion):

Substantial hallucination significantly alters or corrupts the meaning of the paragraph. The substitutions make the text misleading, nonsensical, or factually incorrect.

High risk of misunderstanding, major semantic degradation.

Provide the modified paragraph. Provide the perplexity of the modified paragraph too. Provide a precise yet complete explanation of why the modified paragraph has paragraph-level hallucination and also mention all the sentences in the explanation that you modified from the original paragraph. Follow a common structure for explanation in all the responses. Maintain the following template in the output.
Modified_Paragraph:
Perplexity:
Explanation:

[User Input]:
Title: {**title**}, Abstract: {**abstract**}, Section: {**section**}, Current Paragraph: {**paragraph**}, Previous Paragraph: {**previous_paragraph**}, Next Paragraph: {**next_paragraph**}

Figure 11: Prompt template for generating paragraph-level hallucinated variants.

Non-hallucinated Variant Generation with Token-level Perturbation

[System Input]:
You are given a title, abstract, section and a paragraph of the section from a research paper. Your task is to introduce **non-hallucinated token-level variation** by replacing a few individual content words of the paragraph with related or synonymous words, while keeping the overall meaning of the paragraph intact.

The modified paragraph should be coherent itself and its original relevance with the abstract and title should be preserved.

Do not add or remove entire phrases or sentences. Keep the total number of sentences in the paragraph same.

Only substitute content words (e.g., nouns, verbs, adjectives) with appropriate alternatives that preserve the original meaning.

Do not modify stop words or punctuation.

Do not introduce incorrect, misleading, or fictional information.

Ensure the paragraph remains factually accurate and semantically coherent.

Provide the modified paragraph.
Provide the perplexity of the modified paragraph too.
Provide an explanation of why the substitutions do not distort the meaning and also mention all the words in the explanation that you modified from the original paragraph. Follow a common structure for explanation in all the responses. Maintain the following template in the output.
Modified_Paragraph:
Perplexity:
Explanation:

[User Input]:
Title: {**title**}, Abstract: {**abstract**}, Section: {**section**}, Paragraph: {**paragraph**}

Figure 12: Prompt template for generating non-hallucinated variants with token-level perturbations.

Non-hallucinated Variant Generation with Sentence-level Perturbation

[System Input]:

You are given a title, abstract, section, a current paragraph of the section and its adjacent two paragraphs (previous paragraph and next paragraph) from a research paper. Your task is to introduce **non-hallucinated sentence-level variation** by replacing a few individual sentences of the current paragraph with the paraphrased version of the sentences, while keeping the overall meaning of the current paragraph intact.

The modified current paragraph should be coherent itself and its original relevance with the adjacent paragraphs (if exist), abstract and title should be preserved.

Do not add or remove entire sentences. Keep the total number of sentences in the paragraph same.

Only substitute some sentences with appropriate alternatives that preserve the original meaning.

Do not introduce incorrect, misleading, or fictional information.

Ensure the paragraph remains factually accurate and semantically coherent.

Provide the modified paragraph.

Provide the perplexity of the modified paragraph too.

Provide an explanation of why the substitutions do not distort the meaning and also mention all the sentences in the explanation that you modified from the original paragraph. Follow a common structure for explanation in all the responses. Maintain the following template in the output.

Modified_Paragraph:

Perplexity:

Explanation:

[User Input]:

Title: {**title**}, Abstract: {**abstract**}, Section: {**section**}, Current Paragraph: {**paragraph**}, Previous Paragraph: {**previous_paragraph**}, Next Paragraph: {**next_paragraph**}

Figure 13: Prompt template for generating non-hallucinated variants with sentence-level perturbations.

Original Paragraph
In history, Prince Kornel traveled to Limbangan and moved to Cianjur, then in 1788 was appointed Kepala Cutak (regent assistant) of Cikalong. There are similarities between literature and history, but in the literature, it is deliberately made more detailed. This is to increase the atmosphere and imagination of the reader. Regarding the titimangsa, between the two there is not much difference, which is around the 1780s. In literature, Prince Kornel began to be appointed as regent of Sumedang around 1791. His name changed to Raden Tumenggung Surianagara.

Label: Token-level Hallucination, **Noise Level:** 3

Modified Paragraph 1
In history, Prince Kornel traveled to Limbangan and moved to Cianjur, then in 1788 was appointed Kepala Manta (regent assistant) of Cikalong. There are similarities between literature and history, but in the literature, it is deliberately made more intricate. This is to increase the atmosphere and imagination of the reader. Regarding the titimangsa, between the two there is not much divergence, which is around the 1780s. In literature, Prince Kornel began to be appointed as regent of Sumedang around 1791. His name changed to Raden Tumenggung Luriharta.

Rationale 1
The modified paragraph includes several words that have been replaced to introduce hallucination. "Kepala Cutak" was changed to "Kepala Manta," which alters the meaning of the term. The word "detailed" was replaced with "intricate," which is misleading in this context. The term "difference" was altered to "divergence," which can lead to confusion regarding the comparison being made. Finally, "Surianagara" was replaced with "Luriharta," which is an unrelated and misleading name. Each of these substitutions contributes to a significant distortion of the paragraph's original meaning.

Weighted Score 1: 0.5306

Modified Paragraph 2
In history, Prince Kornel traveled to Limbangan and moved to Cianjur, then in 1788 was appointed Kepala Pupuk (regent assistant) of Cikalong. There are similarities between literature and history, but in the literature, it is deliberately made more detailed. This is to increase the atmosphere and imagination of the reader. Regarding the titimangsa, between the two there is not much difference, which is around the 1780s. In literature, Prince Kornel began to be appointed as regent of Sumedang around 1791. His name changed to Raden Tumenggung Kucinggulung.

Rationale 2
The modified paragraph has introduced token-level hallucination by replacing the words "Cutak" with "Pupuk" and "Surianagara" with "Kucinggulung". These substitutions create a disconnect from the original context, leading to potential confusion regarding historical accuracy. The term "Pupuk" is unrelated to the role of a regent assistant, and "Kucinggulung" does not align with any recognized name or title in the historical context of Prince Kornel. This distorts the meaning and integrity of the paragraph while maintaining the overall structure.

Weighted Score 2: 0.5318

Modified Paragraph 3
In history, Prince Kornel traveled to Limbangan and moved to Cianjur, then in 1988 was appointed Kepala Bantal (regent assistant) of Cikalong. There are similarities between literature and history, but in the literature, it is deliberately made more detailed. This is to increase the atmosphere and imagination of the reader. Regarding the titimangsa, between the two there is not much difference, which is around the 1980s. In literature, Prince Kornel began to be appointed as regent of Sumedang around 1791. His name changed to Raden Tumenggung Surianagara.

Rationale 3
The modified paragraph contains several substantial changes that introduce high levels of hallucination. The year "1788" was replaced with "1988," which is misleading and factually incorrect. The term "Kepala Cutak" was substituted with "Kepala Bantal," which is unrelated and out-of-context in this historical discussion. Additionally, the decade reference changed from "1780s" to "1980s," further distorting the historical accuracy. These substitutions create significant misleading implications about the timeline and roles of Prince Kornel.

Weighted Score 3: 0.5313

Modified Paragraph 4
In history, Prince Kornel traveled to Limbangan and moved to Cianjur, then in 1988 was appointed Kepala Comet (regent assistant) of Cikalong. There are similarities between literature and history, but in the literature, it is deliberately made more detailed. This is to increase the atmosphere and imagination of the reader. Regarding the tambourine, between the two there is not much difference, which is around the 1980s. In literature, Prince Kornel began to be appointed as regent of Sumedang around 1799. His name changed to Raden Tumenggung Cactus.

Rationale 4
The modified paragraph introduces substantial token-level hallucination through several incorrect substitutions. The year "1788" was changed to "1988," which alters the historical context significantly. The term "Kepala Cutak" was replaced with "Kepala Comet," which is nonsensical in this context. The word "titimangsa" was replaced with "tambourine," which is unrelated to the topic. The decade reference changed from "1780s" to "1980s," further distorting the historical accuracy. Additionally, the appointment year for Prince Kornel was altered from "1791" to "1799," and his name was changed from "Raden Tumenggung Surianagara" to "Raden Tumenggung Cactus," introducing unrelated and incorrect entities.

Weighted Score 4: 0.4834 *(minimum)*

Modified Paragraph 5
In history, Prince Kornel traveled to Limbangan and moved to Cianjur, then in 1795 was appointed Kepala Semangka (regent assistant) of Cikalong. There are similarities between literature and history, but in the literature, it is deliberately made more fragrant. This is to increase the atmosphere and imagination of the reader. Regarding the titimangsa, between the two there is not much difference, which is around the 1790s. In literature, Prince Kornel began to be appointed as regent of Sumedang around 1802. His name changed to Raden Tumenggung Surisumar.
Rationale 5
The modified paragraph introduces several words that misalign with the context of the original text. The year "1788" was replaced with "1795," which distorts the timeline. The term "Kepala Cutak" was replaced with "Kepala Semangka," introducing an unrelated concept. The phrase "more detailed" was changed to "more fragrant," which is out of context. The year "1791" was altered to "1802," which further confuses the timeline. Lastly, "Surianagara" was replaced with "Surisumar," creating a misleading named entity. These substitutions significantly impact the clarity and accuracy of the information presented in the paragraph.
Weighted Score 5: 0.5352
Finalized Paragraph ***(Modified Paragraph 4)***
In history, Prince Kornel traveled to Limbangan and moved to Cianjur, then in 1988 was appointed Kepala Comet (regent assistant) of Cikalong. There are similarities between literature and history, but in the literature, it is deliberately made more detailed. This is to increase the atmosphere and imagination of the reader. Regarding the tambourine, between the two there is not much difference, which is around the 1980s. In literature, Prince Kornel began to be appointed as regent of Sumedang around 1799. His name changed to Raden Tumenggung Cactus.
Corresponding Finalized Rationale
The modified paragraph introduces substantial token-level hallucination through several incorrect substitutions. The year "1788" was changed to "1988," which alters the historical context significantly. The term "Kepala Cutak" was replaced with "Kepala Comet," which is nonsensical in this context. The word "titimangsa" was replaced with "tambourine," which is unrelated to the topic. The decade reference changed from "1780s" to "1980s," further distorting the historical accuracy. Additionally, the appointment year for Prince Kornel was altered from "1791" to "1799," and his name was changed from "Raden Tumenggung Surianagara" to "Raden Tumenggung Cactus," introducing unrelated and incorrect entities.

Table 7: Example of token-level hallucinated variants and associated rationales.

Original Paragraph
Findings of the research were classified into some patterns with regard to students' problems observed in the post-editing process. There were seven distinctive problems found in their translation practices, i.e. terminological problems, inappropriate use of grammatical rules, irrelevant word choices, inconsistent use of technical terms, violations in translating collocations, inadequate subject knowledge of the translated texts and uncontrolled source language texts. Details of each finding are described in the following sections.

Label: Sentence-level Hallucination, **Noise Level:** 1

Modified Paragraph 1
Findings of the research were classified into some patterns with regard to students' problems observed in the post-editing process. There were seven distinctive problems found in their translation practices, i.e. terminological problems, inappropriate use of grammatical rules, irrelevant word choices, the use of abstract concepts in practical applications, violations in translating collocations, inadequate subject knowledge of the translated texts and uncontrolled source language texts. Details of each finding are described in the following sections.

Rationale 1
The modified paragraph retains the overall structure and theme but introduces sentence-level hallucination by replacing the sentence "inconsistent use of technical terms" with "the use of abstract concepts in practical applications." This replacement introduces a logical inconsistency, as it shifts the focus from a specific issue in translation to a more general and unrelated concept. This creates a slight disconnect in the context of the paragraph, maintaining coherence in theme but distorting local context.

Weighted Score 1: 0.6027

Modified Paragraph 2
Findings of the research were classified into some patterns with regard to students' problems observed in the post-editing process. There were seven distinctive problems found in their translation practices, i.e. the impact of social media on language acquisition, irrelevant word choices, inconsistent use of technical terms, violations in translating collocations, inadequate subject knowledge of the translated texts and uncontrolled source language texts. Details of each finding are described in the following sections.

Rationale 2
The modified paragraph introduces sentence-level hallucination by replacing the first sentence of the current paragraph with a completely unrelated statement about the impact of social media on language acquisition. This substitution disrupts the logical flow of the paragraph while still maintaining the overall theme of discussing students' problems in translation practices. The original sentence was "There were seven distinctive problems found in their translation practices, i.e. terminological problems, inappropriate use of grammatical rules," which has been altered to include a false claim. The rest of the sentences remain unchanged to preserve overall coherence.

Weighted Score 2: 0.5709 *(minimum)*

Modified Paragraph 3
Findings of the research were classified into some patterns with regard to students' problems observed in the post-editing process. There were seven distinctive problems found in their translation practices, i.e. terminological problems, the overuse of metaphors in academic writing, irrelevant word choices, inconsistent use of technical terms, violations in translating collocations, inadequate subject knowledge of the translated texts and uncontrolled source language texts. Details of each finding are described in the following sections.

Rationale 3
The modified paragraph contains sentence-level hallucination primarily in the second sentence. The original sentence stated, "inappropriate use of grammatical rules," which was replaced with "the overuse of metaphors in academic writing." This substitution introduces a logical inconsistency as it does not align with the context of translation problems. The other sentences remain unchanged, maintaining the overall structure but impacting local coherence.

Weighted Score 3: 0.5861

Modified Paragraph 4
Findings of the research were classified into some patterns with regard to students' problems observed in the post-editing process. There were seven distinctive problems found in their translation practices, i.e. the influence of weather patterns on linguistic choices, inappropriate use of grammatical rules, irrelevant word choices, inconsistent use of technical terms, violations in translating collocations, inadequate subject knowledge of the translated texts and uncontrolled source language texts. Details of each finding are described in the following sections.

Rationale 4
The modified paragraph introduces hallucination by replacing the first sentence with an unrelated sentence regarding weather patterns, which is not relevant to the context of post-editing in translation practices. This substitution alters the logical flow of the paragraph while maintaining the overall theme of discussing student problems in translation. The sentence replaced is: "There were seven distinctive problems found in their translation practices, i.e. terminological problems." The modified paragraph retains coherence but introduces a misleading element that detracts from the clarity of the findings.

Weighted Score 4: 0.5758

Modified Paragraph 5

Findings of the research were classified into some patterns with regard to students' problems observed in the post-editing process. There were seven distinctive problems found in their translation practices, i.e. an overuse of punctuation marks, incorrect sentence structures, irrelevant word choices, inconsistent use of technical terms, frequent shifts in narrative perspective, inadequate subject knowledge of the translated texts and uncontrolled source language texts. Details of each finding are described in the following sections.
Rationale 5
The modified paragraph introduces sentence-level hallucination by replacing the original sentences with unrelated content that still loosely fits the theme of translation problems. Specifically, I modified the second sentence. The original sentence discussed "terminological problems" and "inappropriate use of grammatical rules," which were replaced with "an overuse of punctuation marks" and "incorrect sentence structures." This introduces logical inconsistency regarding the types of problems faced by the students, while still addressing the theme of translation difficulties.
Weighted Score 5: 0.5808
Finalized Paragraph ***(Modified Paragraph 2)***
Findings of the research were classified into some patterns with regard to students' problems observed in the post-editing process. There were seven distinctive problems found in their translation practices, i.e. the impact of social media on language acquisition, irrelevant word choices, inconsistent use of technical terms, violations in translating collocations, inadequate subject knowledge of the translated texts and uncontrolled source language texts. Details of each finding are described in the following sections.
Corresponding Finalized Rationale
The modified paragraph introduces sentence-level hallucination by replacing the first sentence of the current paragraph with a completely unrelated statement about the impact of social media on language acquisition. This substitution disrupts the logical flow of the paragraph while still maintaining the overall theme of discussing students' problems in translation practices. The original sentence was "There were seven distinctive problems found in their translation practices, i.e. terminological problems, inappropriate use of grammatical rules," which has been altered to include a false claim. The rest of the sentences remain unchanged to preserve overall coherence.

Table 8: Example of sentence-level hallucinated variants and associated rationales.

Original Paragraph

Regarding the player's study period, it can be seen that player AB has studied Japanese for three years in her school extracurricular. Player AB learns hiragana, katakana, basic grammar, and beginner level of kanji. After graduating from high school in 2019, Player AB has not continued to study Japanese anymore. Then in 2020, player AB started playing the A3! visual novel game, a Japanese language game. Even though player AB only learned basic Japanese, it can be seen that Player AB has a high understanding of vocabulary, and there are only a few words that she does not know.

Label: Paragraph-level Hallucination, **Noise Level:** 3

Modified Paragraph 1

Regarding the player's culinary experience, it can be seen that player AB has been involved in gourmet cooking for three years as part of her culinary extracurricular activities. Player AB learns various cooking techniques, flavor profiles, and beginner-level gastronomy. After graduating from high school in 2019, Player AB has not continued her culinary education anymore. Then in 2020, player AB started participating in a cooking competition, a culinary-focused event. Even though player AB only learned basic cooking skills, it can be seen that Player AB has a high understanding of flavor combinations, and there are only a few techniques that she does not know.

Rationale 1

The modified paragraph has undergone significant changes that deviate from the original context of Japanese language learning through the A3! visual novel game. The main topic has shifted from the study of Japanese language skills to culinary experiences, which is entirely unrelated to the overall theme of the paper. Specific modifications include:
1. "Regarding the player's study period" was changed to "Regarding the player's culinary experience." 2. "studied Japanese" was changed to "been involved in gourmet cooking" 3. "extracurricular" was changed to "culinary extracurricular activities." 4. "Player AB learns hiragana, katakana, basic grammar, and beginner level of kanji" was altered to "Player AB learns various cooking techniques, flavor profiles, and beginner-level gastronomy." 5. "continued to study Japanese anymore" was modified to "continued her culinary education anymore." 6. "A3! visual novel game, a Japanese language game" was replaced with "a cooking competition, a culinary-focused event." 7. "Japanese" was replaced with "cooking skills" 8. "understanding of vocabulary" was changed to "understanding of flavor combinations." 9. "words that she does not know" became "techniques that she does not know."
These changes introduced a completely new theme centered around cooking, making the paragraph inconsistent with the title, abstract, and surrounding paragraphs that focus on Japanese language learning.

Weighted Score 1: 0.2412 *(minimum)*

Modified Paragraph 2

Regarding the player's study period, it can be seen that player AB has studied culinary arts for three years in her school extracurricular. Player AB learns various cooking techniques, basic nutrition, and beginner level of food presentation. After graduating from high school in 2019, Player AB has not continued to study culinary arts anymore. Then in 2020, player AB started exploring international cuisines, particularly focusing on Italian cooking. Even though player AB only learned basic culinary skills, it can be seen that Player AB has a high understanding of flavors, and there are only a few techniques that she does not know.

Rationale 2

The modified paragraph introduces a completely different theme focused on culinary arts instead of Japanese language skills, which deviates significantly from the overall context of the research on the A3! visual novel game. The original paragraph discussed Player AB's experience with studying Japanese, her learning of language components, and the influence of the A3! game on her vocabulary comprehension. The modified sentences are: 1. "Player AB has studied culinary arts for three years in her school extracurricular." 2. "Player AB learns various cooking techniques, basic nutrition, and beginner level of food presentation." 3. "Then in 2020, player AB started exploring international cuisines, particularly focusing on Italian cooking." 4. "Even though player AB only learned basic culinary skills, it can be seen that Player AB has a high understanding of flavors, and there are only a few techniques that she does not know."
These changes completely shift the focus from language learning to cooking, making the paragraph misleading and inconsistent with the title and abstract.

Weighted Score 2: 0.2994

Modified Paragraph 3

Regarding the player's culinary interests, it can be seen that player AB has studied various cuisines during her travels abroad. Player AB learns the intricacies of French, Italian, and traditional Japanese cooking methods. After returning from her culinary studies in 2019, Player AB has not pursued professional cooking anymore. Then in 2020, player AB started exploring the world of fusion cuisine, a blend of different culinary traditions. Even though player AB only learned basic cooking techniques, it can be seen that Player AB has a high understanding of flavor profiles, and there are only a few ingredients that she does not know.

Rationale 3

The modified paragraph introduces a completely different theme focused on culinary interests rather than the influence of the A3! visual novel game on Japanese language skills. The sentences were altered to change the context from learning a language to exploring cooking, making it misleading and inconsistent with the paper's focus on language learning through a visual novel game. The original sentences regarding the study of Japanese were replaced with sentences about studying various cuisines, and the references to Japanese language skills were changed to understanding flavor profiles. The sentences modified include: 1. "Regarding the player's study period" → "Regarding the player's culinary interests" 2. "has studied Japanese for three years in her school extracurricular" → "has studied various cuisines during her travels abroad." 3. "learns hiragana, katakana, basic grammar, and beginner level of kanji" → "learns the intricacies of French, Italian, and traditional Japanese cooking methods." 4. "After graduating from high school in 2019, Player AB has not continued to study Japanese anymore" → "After returning from her culinary studies in 2019, Player AB has not pursued professional cooking anymore." 5. "Then in 2020, player AB started playing the A3! visual novel game, a Japanese language game" → "Then in 2020, player AB started exploring the world of fusion cuisine, a blend of different culinary traditions." 6. "Even though player AB only learned basic Japanese" → "Even though player AB only learned basic cooking techniques." 7. "high understanding of vocabulary" → "high understanding of flavor profiles." 8. "a few words that she does not know" → "a few ingredients that she does not know."

Weighted Score 3: 0.2912

Modified Paragraph 4

Regarding the player's study period, it can be seen that player AB has studied the art of origami for three years in her school extracurricular. Player AB learns various folding techniques, basic patterns, and beginner level of paper types. After graduating from high school in 2019, Player AB has not continued to explore origami anymore. Then in 2020, player AB started creating her own origami projects, a traditional Japanese art form. Even though player AB only learned basic origami, it can be seen that Player AB has a high understanding of complex designs, and there are only a few techniques that she does not know.

Rationale 4

The modified paragraph has deviated significantly from the context of the original paper, which focused on the influence of the A3! visual novel game on players' Japanese language skills. The main topic has shifted from Japanese language learning to the art of origami, which is unrelated to the research theme.

1. "player AB has studied Japanese for three years in her school extracurricular" was changed to "player AB has studied the art of origami for three years in her school extracurricular." 2. "Player AB learns hiragana, katakana, basic grammar, and beginner level of kanji" was changed to "Player AB learns various folding techniques, basic patterns, and beginner level of paper types." 3. "After graduating from high school in 2019, Player AB has not continued to study Japanese anymore" was changed to "After graduating from high school in 2019, Player AB has not continued to explore origami anymore." 4. "Then in 2020, player AB started playing the A3! visual novel game, a Japanese language game" was changed to "Then in 2020, player AB started creating her own origami projects, a traditional Japanese art form." 5. "Even though player AB only learned basic Japanese, in Figure 4, it can be seen that Player AB has a high understanding of vocabulary, and there are only a few words that she does not know" was changed to "Even though player AB only learned basic origami, it can be seen that Player AB has a high understanding of complex designs, and there are only a few techniques that she does not know."

Weighted Score 4: 0.3428

Modified Paragraph 5

Regarding the player's culinary preferences, it can be seen that player AB has developed a keen interest in Italian cuisine through her school extracurricular activities. Player AB learns about pasta, sauces, basic cooking techniques, and beginner-level baking. After graduating from high school in 2019, Player AB has not continued to pursue her culinary education anymore. Then in 2020, player AB started exploring various Italian recipes, a culinary adventure. Even though player AB only learned basic cooking skills, it can be seen that Player AB has a high understanding of flavor combinations, and there are only a few techniques that she does not know.

Rationale 5

The modified paragraph introduces a completely new theme centered on culinary preferences and cooking education, which deviates significantly from the original context of discussing a player's experience with learning Japanese through the A3! visual novel game. The key changes made include:

1. "Regarding the player's study period" was changed to "Regarding the player's culinary preferences." 2. "has studied Japanese for three years in her school extracurricular" was modified to "has developed a keen interest in Italian cuisine through her school extracurricular activities." 3. "Player AB learns hiragana, katakana, basic grammar, and beginner level of kanji" was altered to "Player AB learns about pasta, sauces, basic cooking techniques, and beginner-level baking." 4. "Player AB started playing the A3! visual novel game, a Japanese language game" became "player AB started exploring various Italian recipes, a culinary adventure." 5. "understanding of vocabulary" was changed to "understanding of flavor combinations." 6. "a few words that she does not know" was modified to "a few techniques that she does not know."
These alterations create a substantial semantic drift, shifting the focus from language learning to culinary interests, leading to a high risk of misunderstanding.

Weighted Score 5: 0.2843

Finalized Paragraph ***(Modified Paragraph 1)***

Regarding the player's culinary experience, it can be seen that player AB has been involved in gourmet cooking for three years as part of her culinary extracurricular activities. Player AB learns various cooking techniques, flavor profiles, and beginner-level gastronomy. After graduating from high school in 2019, Player AB has not continued her culinary education anymore. Then in 2020, player AB started participating in a cooking competition, a culinary-focused event. Even though player AB only learned basic cooking skills, it can be seen that Player AB has a high understanding of flavor combinations, and there are only a few techniques that she does not know.
Corresponding Finalized Rationale
The modified paragraph has undergone significant changes that deviate from the original context of Japanese language learning through the A3! visual novel game. The main topic has shifted from the study of Japanese language skills to culinary experiences, which is entirely unrelated to the overall theme of the paper. Specific modifications include: 1. "Regarding the player's study period" was changed to "Regarding the player's culinary experience." 2. "extracurricular" was changed to "culinary extracurricular activities." 3. "Player AB learns hiragana, katakana, basic grammar, and beginner level of kanji" was altered to "Player AB learns various cooking techniques, flavor profiles, and beginner-level gastronomy." 4. "continued to study Japanese anymore" was modified to "continued her culinary education anymore." 5. "A3! visual novel game, a Japanese language game" was replaced with "a cooking competition, a culinary-focused event." 6. "understanding of vocabulary" was changed to "understanding of flavor combinations." 7. "words that she does not know" became "techniques that she does not know." These changes introduced a completely new theme centered around cooking, making the paragraph inconsistent with the title, abstract, and surrounding paragraphs that focus on Japanese language learning.

Table 9: Example of paragraph-level hallucinated variants and associated rationales.

Original Paragraph

Nyimur tradition is likened to health immunization in modern society. Based on the knowledge they know, with this tradition, their children become safer to live their own lives in the future. Nyimur tradition is a ceremony that should be followed by all residents who have got a baby in KasepuhanCiptagelar.

Label: No Hallucination, **Noise Level:** 0

Modified Paragraph 1

Nyimur tradition is compared to health vaccination in contemporary society. Based on the understanding they possess, with this tradition, their children become more secure to lead their own lives in the future. Nyimur tradition is a ritual that should be observed by all inhabitants who have welcomed a baby in Kasepuhan Ciptagelar.

Rationale 1

The substitutions made in the paragraph were: "likened" to "compared," "immunization" to "vaccination," "knowledge" to "understanding," "safer" to "more secure," "live" to "lead," "ceremony" to "ritual," "followed" to "observed," and "residents" to "inhabitants." These changes maintain the original meaning of the paragraph by conveying the same ideas about the Nyimur tradition and its importance in the community while using different but related terminology. The overall context regarding the tradition's role in promoting safety and its observance among community members remains intact.

Weighted Score 1: 0.8639

Modified Paragraph 2

Nyimur tradition is compared to health immunization in contemporary society. Based on the understanding they possess, with this tradition, their children become more secure to live their own lives in the future. Nyimur tradition is a ritual that should be observed by all inhabitants who have welcomed a baby in Kasepuhan Ciptagelar.

Rationale 2

The substitutions made in the modified paragraph include "likened" to "compared," "modern" to "contemporary," "knowledge" to "understanding," "safer" to "more secure," "ceremony" to "ritual," and "residents" to "inhabitants," as well as "got" to "welcomed." These changes do not distort the meaning of the original text as they maintain the same context and convey the same message about the significance of the Nyimur tradition in relation to child safety and community practices. Each replacement word is contextually appropriate and synonymous, ensuring that the overall essence and coherence of the paragraph are preserved.

Weighted Score 2: 0.8782

Modified Paragraph 3

Nyimur tradition is compared to health immunization in contemporary society. Based on the wisdom they possess, with this tradition, their children become more secure to lead their own lives in the future. Nyimur tradition is a ritual that should be observed by all inhabitants who have welcomed a baby in Kasepuhan Ciptagelar.

Rationale 3

The substitutions made include "likened" to "compared," "modern" to "contemporary," "knowledge" to "wisdom," "know" to "possess," "safer" to "more secure," "live" to "lead," "ceremony" to "ritual," and "residents" to "inhabitants," which all maintain the original meaning of the paragraph. These modifications do not distort the meaning as they convey similar concepts and maintain the context of the Nyimur tradition's significance in Kasepuhan Ciptagelar. The overall coherence and relevance to the title and abstract are preserved.

Weighted Score 3: 0.8793 *(maximum)*

Modified Paragraph 4

Nyimur tradition is compared to health vaccination in contemporary society. Based on the understanding they possess, with this tradition, their children become more secure to lead their own lives in the future. Nyimur tradition is a ritual that should be observed by all inhabitants who have welcomed a baby in Kasepuhan Ciptagelar.

Rationale 4

The modifications made in the paragraph include the following substitutions: "likened" to "compared," "immunization" to "vaccination," "modern" to "contemporary," "knowledge" to "understanding," "safer" to "more secure," "live" to "lead," "ceremony" to "ritual," and "residents" to "inhabitants." These changes do not distort the meaning of the original paragraph as they maintain the core concepts and relationships present. The essence of the discussion regarding the Nyimur tradition and its significance for the community remains intact.

Weighted Score 4: 0.8628

Modified Paragraph 5

Nyimur tradition is compared to health immunization in contemporary society. Based on the knowledge they possess, with this tradition, their children become more secure to live their own lives in the future. Nyimur tradition is a ritual that should be observed by all inhabitants who have welcomed a baby in Kasepuhan Ciptagelar.

Rationale 5

The modifications made in the paragraph include replacing "likened" with "compared," "modern" with "contemporary," "know" with "possess," "safer" with "more secure," "ceremony" with "ritual," and "residents" with "inhabitants." These substitutions maintain the overall meaning of the original paragraph while introducing slight variations in word choice. The essence of the content remains intact, ensuring that the focus on the Nyimur tradition and its significance for the community of Kasepuhan Ciptagelar is preserved.
Weighted Score 5: 0.8757
Finalized Paragraph ***(Modified Paragraph 3)***
Nyimur tradition is compared to health immunization in contemporary society. Based on the wisdom they possess, with this tradition, their children become more secure to lead their own lives in the future. Nyimur tradition is a ritual that should be observed by all inhabitants who have welcomed a baby in Kasepuhan Ciptagelar.
Corresponding Finalized Rationale
The substitutions made include "likened" to "compared," "modern" to "contemporary," "knowledge" to "wisdom," "know" to "possess," "safer" to "more secure," "live" to "lead," "ceremony" to "ritual," and "residents" to "inhabitants," which all maintain the original meaning of the paragraph. These modifications do not distort the meaning as they convey similar concepts and maintain the context of the Nyimur tradition's significance in Kasepuhan Ciptagelar. The overall coherence and relevance to the title and abstract are preserved.

Table 10: Example of non-hallucinated variants and associated rationales.