Overview of the SciHal25 Shared Task on Hallucination Detection for Scientific Content

Dan Li **Bogdan Palfi** **Colin Kehang Zhang** **Jaiganesh Subramanian**
Adrian Raudaschl **Yoshiko Kakita**
Anita De Waard **Zubair Afzal** **Georgios Tsatsaronis**
Elsevier
{d.li1, b.palfi, c.zhang.3, j.subramanian1, a.raudaschl, y.kakita,
a.dewaard, zubair.afzal, g.tsatsaronis}@elsevier.com

Abstract

This paper provides an overview of the Hallucination Detection for Scientific Content (SciHal) shared task held in the 2025 ACL Scholarly Document Processing workshop. The task invites participants to detect hallucinated claims in answers to research-oriented questions generated by real-world GenAI-powered research assistants. This task is formulated as a multi-label classification problem, each instance consists of a question, an answer, an extracted claim, and supporting reference abstracts. Participants are asked to label claims under two subtasks: (1) coarse-grained detection with labels Entailment, Contradiction, or Unverifiable; and (2) fine-grained detection with a more detailed taxonomy including 8 types. The dataset consists of 488 research-oriented questions collected over one week from a generative assistant tool. These questions were rewritten using GPT-4o and manually reviewed to address potential privacy or commercial concerns. In total, approximately 10,000 reference abstracts were retrieved, and 4,592 claims were extracted from the assistant's answers. Each claim is annotated with hallucination labels. The dataset is divided into 3,592 training, 500 validation, and 500 test instances. Subtask 1 saw 109 submissions across 11 teams while subtask 2 saw 43 submissions across 7 teams, resulting in a total of 5 published technical reports. This paper summarizes the task design, dataset, participation, and key findings.

1 Introduction

Generative AI-powered academic research assistants are transforming how research is conducted. These systems enable users to pose research-related questions in natural language and receive structured, concise summaries supported by relevant references. However, hallucinations pose a significant challenge to fully trusting these automatically generated scientific answers.

Recent shared tasks have begun to address hallucination detection across domains such as biomedical summarization (Gupta et al., 2024) and scientific content (Mickus et al., 2024). While these efforts have advanced benchmarking in specific settings, they are often limited to binary classification or constrained domains. Broader benchmarks like Hal-Eval (Jiang et al., 2024) provide general-purpose evaluation but lack task grounding.

To fill this gap, SciHal introduces a multi-label hallucination detection task grounded in real-world scientific question answering. The task invites participants to detect hallucinated claims in answers to research-oriented questions generated by a real-world GenAI-powered research assistant. This task is formulated as a multi-label classification problem, each instance consists of a question, an answer, an extracted claim, and supporting reference abstracts. The shared task is hosted on Kaggle[1].

Weighted F1 score is used as the primary evaluation metric to account for class imbalance. Subtask 1 attracted 109 submissions from 11 participating teams. On the public leaderboard (validation set), the top three teams were Schopf et al. (2025), Cao et al. (2025), and Le and Thin (2025), achieving weighted F1 scores of 0.60, 0.59, and 0.59, respectively. On the private leaderboard (test set), the top three teams were Schopf et al. (2025), Cao et al. (2025), and Galimzianova et al. (2025), with scores of 0.62, 0.60, and 0.59.

Subtask 2 attracted 43 submissions from 7 participating teams. On the public leaderboard (validation set), the top three teams were Cao et al. (2025), Schopf et al. (2025), and JB, achieving weighted F1 scores of 0.51, 0.50, and 0.49. On the private leaderboard (test set), the top teams were Schopf et al. (2025), Cao et al. (2025), Le and Thin (2025), JB, Carla and Uban (2025), achieving weighted F1

[1] https://www.kaggle.com/competitions/hallucination-detection-scientific-content-2025

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 307–315

scores of 0.47, 0.47, 0.47, 0.47, 0.46.

These results highlight the difficulty and complexity of the task. Participating teams employed a diverse range of approaches, including fine-tuning transformer-based encoders, prompting large language models (LLMs), and hybrid methods using internal state representations. Additionally, the subjective nature of hallucination detection, particularly in edge cases, introduces annotation challenges and potential label noise. Improving annotation consistency remains an important direction for future work.

2 Related Work

Recent years have seen a growing interest in shared tasks on hallucination detection in the context of text generation by LLMs. One of the earliest domain-specific efforts is the TREC BioGen task, which evaluates the factual consistency of biomedical answers and summaries, using sentence-level labels over content generated from PubMed articles (Gupta et al., 2024). In the scientific domain, the SHROOM shared task (Mickus et al., 2024) introduced hallucination detection and mitigation challenges for scientific abstracts and question answering, incorporating both binary and fine-grained classifications. The SHROOM dataset includes human-annotated claims with hallucination labels grounded in scientific references, offering valuable insights but remaining limited in scale and question diversity. Beyond biomedical and scientific settings, the Hal-Eval benchmark (Jiang et al., 2024) provides a multi-domain benchmark covering summarization, question answering, and data-to-text generation, annotated with fine-grained hallucination spans.

Although these efforts contribute valuable datasets and evaluation protocols, they often focus on either general-purpose outputs, a single domain, or a binary classification setup. In contrast, SciHal is specifically designed for hallucination detection in academic research assistants. It introduces a two-tiered taxonomy (coarse- and fine-grained), grounded in real-world user queries and scientific reference abstracts, with large-scale expert annotations across five scientific domains. This makes SciHal the first shared task to target hallucination detection in the context of retrieval-augmented question answering for scholarly research.

3 Hallucination Taxonomy Creation

There is currently no established taxonomy for hallucination types specific to scientific content. Our goal is to develop one that (1) reflects real-world error patterns, (2) remains manageable for human annotators, and (3) ensures high label quality.

Existing work has developed detailed taxonomies to characterize hallucinations in large language models (LLMs). Early studies often framed hallucinations as a binary phenomenon, i.e. factual versus non-factual, but more recent work proposes nuanced classifications. A common distinction is between *intrinsic* hallucinations, which contradict the input or reference, and *extrinsic* hallucinations, which introduce unsupported content (Huang et al., 2023; Zhang et al., 2023). Other taxonomies categorize hallucinations based on the nature of the error, such as entity-level, numeric, or reasoning-based inconsistencies (Mishra et al., 2024; Li et al., 2024). Some frameworks adopt a multi-dimensional view; for example, Rawte et al. (2023) organize hallucinations by orientation (harmful vs. benign), grounding (intrinsic vs. extrinsic), and fine-grained types, including acronym misuse, quantitative errors, and temporal inaccuracies. These efforts provide a foundation for designing task-specific taxonomies in domains like scientific content generation (Hu et al., 2024).

Drawing from a small-scale analysis of 136 user feedback responses, we identified the most frequent error types: missing the main concept (34.6%), factually incorrect (21.3%), too general (21.3%), and unrelated references (11.8%). These findings highlight recurring issues in generative AI outputs.

Our final taxonomy is informed by both in-house analysis of GenAI-powered research assistant outputs and broader studies of hallucination patterns in general-purpose GenAI systems. Figure 1 presents the decision tree that underpins our taxonomy, which was included in the annotation guidelines provided to subject-matter experts (SMEs). Definitions and examples for each hallucination type are listed in Table 1.

4 Data Creation

The dataset consists of claim-level annotations designed to evaluate the factual consistency between claims in generated answers and their cited references within scientific retrieval-augmented generation (RAG) systems. The data are primarily derived

T1 label	T2 label	Definition	Examples
entail	entail	The claim is explicitly and clearly supported by at least one passage in the reference abstracts, while not being contradicted by any other passage from the reference.	*Reference: The weather is rainy and the wind is blowing. Claim: The weather is rainy.*
unver	unrelunvef	The claim and the abstracts address different topics, therefore making the claim unverifiable.	*Reference: The weather is rainy and the wind is blowing. Claim 1: He was born in the Netherlands. → The reference addresses the weather but the claim mentions where a person was born, being unrelated so unverifiable.*
unver	relunvef	The claim and all abstracts address the same broad topic, but the specific idea presented in the claim or any of its sub-parts is not covered, making the claim unverifiable.	*Reference: The weather is rainy and the wind is blowing. Claim 1: The rainy weather is causing widespread flooding in the region. → Both the reference and the claim address the weather, but nothing is mentioned about flooding.*
contra	entierr	The claim contains an erroneous entity that contradicts what is stated in the reference. A named entity is a real-world object, such as a person, location, organization, product, etc., that can be denoted with a proper name.	*Reference: The weather is rainy, as forecasted by BBC. Claim: The weather is rainy, as forecasted by The Weather Channel.*
contra	numerr	The claim contains an erroneous numeric value that contradicts the reference.	*Reference: The concentration was 80%. Claim: The concentration was 90%.*
contra	negat	The claim negates parts of the reference or replaces terms with their antonyms, therefore stating the opposite to what appears in the reference.	*Reference: It is windy and the temperature is increasing. Claim: It is not windy and the temperature is decreasing.*
contra	missinfo	The claim omits critical information from the reference, leading to an incorrect or incomplete understanding of the reference. This can occur when the reference abstract makes a conditional statement like: "when / if / by X then Y", but the condition is missing.	*Reference: Regular exercise, when performed consistently and in combination with a balanced diet and healthy lifestyle, can reduce the risk of heart disease by 30% and also improve mental health. Claim: Regular exercise mainly enhances mental well-being. → Missing critical info: omits the condition "when performed consistently and in combination with... ."*
contra	misinter	The claim presents logical fallacies, flawed reasoning or illogical conclusions through overclaiming, under-claiming, ambiguity, inconsistency or implying a consensus among references when there are disagreements.	*Reference: Regular exercise can reduce the risk of heart disease by 30% and also improve mental health. Claim 1: Regular exercise eliminates the risk of heart disease. → Overstatement. Claim 2: Only regular exercise is required for improved mental health. → Logical fallacy.*

Table 1: The definitions of hallucination types.

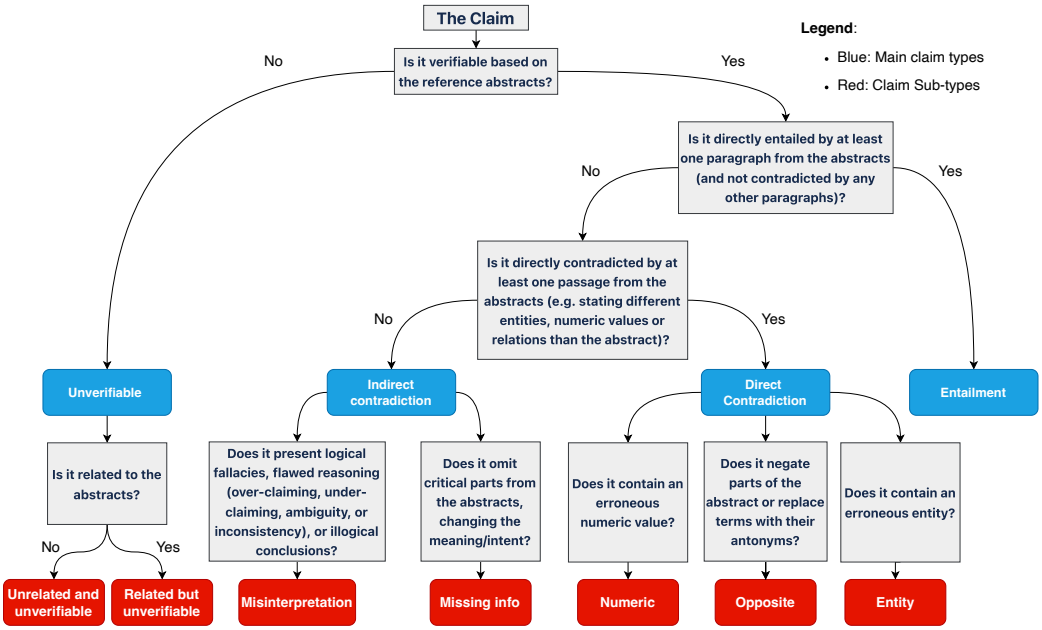

Figure 1: The hallucination taxonomy.

from Scopus AI[2], an in-house research assistant tool powered by a RAG system indexing millions of scientific abstracts.

4.1 Question, Answer, and Claim Collection

We first collected over 50,000 real-user questions from Scopus AI. Using a large language model (LLM), we classified each question by domain and verified its correctness, completeness, and language. Only English questions that were correct and complete were retained. Considering the popularity and availability of SMEs, we keep questions in the five domains – Engineering, Environmental Science, Medicine, Agricultural and Biological Sciences, and Computer Science. We then used an LLM to rewrite the queries, manually spot-checking them to remove privacy or commercial concerns, resulting in 500 questions for hallucination label annotation.

Next, we used the Scopus AI endpoint service to generate answers for the questions. Each answer was supported by up to 20 reference abstracts. We then extracted claims from each answer along with their corresponding references.

4.2 Inducing Hallucinations into Claims

To balance the class distribution, we introduced synthetic hallucinations into the claims via LLM prompting. An in-house annotator processed the original data, and 65% of the claims were randomly selected for modification, where an LLM induced hallucinations based on predefined types (Sub-task 2) while maintaining type balance. The dataset, comprising 35% original claims and 65% error-induced claims, was then sent to subject matter experts (SMEs) for annotation. This approach allows us to estimate the hallucination rate of the in-house research assistant using the original claims while ensuring a balanced dataset, where entailment accounts for less than 35% and other types each account for under 10%.

4.3 SME Annotation Process

SME annotation was conducted via external vendors. Annotators were provided with the data to be labeled, including the question, generated answer, extracted claim, and list of reference abstracts, as well as detailed annotation guidelines. These guidelines included definitions of hallucination types and a decision tree to support consistent labeling. A trial phase was conducted to ensure alignment with the guidelines before full-scale annotation.

To balance annotation quality and cost, we adopted a hybrid strategy that combined human SME labels with predictions from an internal LLM-based hallucination detection model. In the initial annotation phase, each instance was labeled by one domain-specific SME, who provided both a

[2] https://www.elsevier.com/products/scopus/scopus-ai

hallucination label and a brief textual justification (1 – 3 sentences). We then compared the SME-provided label with the LLM-generated prediction. Instances where both sources agreed were grouped into Batches 1 and 2, which we consider to be consistent and cost-effective, as they rely on a single SME confirmation.[3]

In the final annotation phase, instances where the SME and LLM disagreed were re-labeled by a second SME. A third SME then adjudicated, having access to both prior labels and their justifications. This adjudication step ensures high-quality, consensus-based annotations. The resulting data were split into Batch 3 (training), validation, and test sets. These subsets are particularly valuable: they are both challenging as they are derived from disagreement cases between humans and LLM, and they are reliable as they reflect consensus among two or three SMEs. Note that a few invalid instances were removed and this resulted in 488 questions in the final release data.

5 Analysis of Label Quality

The following sections will present an analysis of the label distribution and quality. To this end, it is important to note that a second SME was consulted only for claims where there was disagreement between the LLM judge and the first SME. As a result, the comparisons between SME 1 and SME 2 in the following analysis relate specifically to claims that are potentially more difficult to label or more subjective. These challenging claims constitute approximately 50% of the entire dataset. Consequently, the observed agreement rate between the two SMEs may be lower than if the comparison were conducted on the entire dataset. The third SME was excluded from this analysis because they were not independent, having had access to both the labels and justifications provided by the first two SMEs.

Overall, the agreement rates (accuracy) between the two SMEs on the **difficult** part of the entire dataset are **0.55** for Subtask 1 and **0.44** for Subtask 2, respectively.

5.1 Subtask 1

Table 2 shows that subtask 1 exhibits a relatively balanced distribution of labels, with Entailment and Contradiction accounting for approximately 38.71% and 36.89%, respectively, while Unverifiable claims are less frequent at 24.4%. Comparing the agreement between the two SMEs, Figure 2 and Table 3 reveal that the SMEs tend to agree more often when labeling claims as Entailment, while showing the highest disagreement when classifying Unverifiable claims. This may suggest that some SMEs are more strict when assigning the Entailment label. Overall, the agreement rate (accuracy) is 0.55, indicating that the SMEs concur in more than half of the cases. The Cohen's Kappa coefficient of 0.297 further reflects this trend, signifying a fair level of agreement where disagreements still occur between the raters. These results highlight the necessity of, and motivate our decision to, involve a third SME to adjudicate disagreements and aggregate the labels, thereby ensuring higher data quality.

Label	Count	Percentage
entail	1762	38.71%
unver	1111	24.40%
contra	1677	36.89%

Table 2: Claim distribution for the full Subtask 1 dataset.

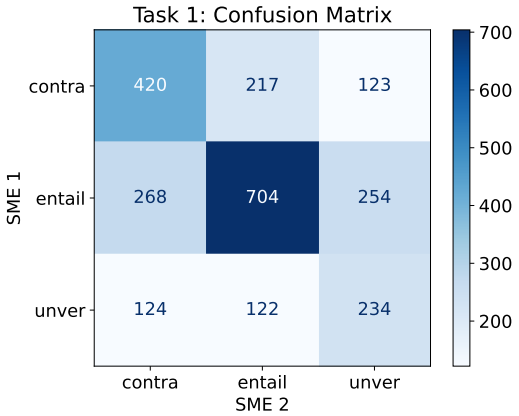

Figure 2: Confusion matrix comparing the predictions of two independent SMEs for Subtask 1. This figure is based solely on a more challenging subset of the data, comprising approximately 50% of the entire dataset, for which labels from the second SME are available.

5.2 Subtask 2

The second task involves a finer-grained classification, further subdividing the unverifiable and contradicted claims into multiple sub-types. The distribution of these sub-types is presented in Table 4. Notably, the majority of unverifiable claims

[3]Batch 1 is a subset of Batch 2 and will be deprecated in future updates. We recommend using Batches 2 and 3 for training.

	Precision	Recall	F1	Support
contra	0.55	0.52	0.53	812
entail	0.57	0.67	0.62	1043
unver	0.49	0.38	0.43	611

Table 3: Classification report comparing the predictions of two independent SMEs for Subtask 1. This report is based solely on a more challenging subset of the data, comprising approximately 50% of the entire dataset, for which labels from the second SME are available.

are related to the reference, comprising approximately 20.34% of the total, whereas only 4.07% are unrelated. Among contradicted claims, the most frequent sub-types are negations or opposite statements (15.4%) and misinterpretations (10.83%), while the remaining sub-types each account for less than 6% of cases.

Inter-annotator agreement between the two SMEs is showed in Figure 3 and Table 5. The SMEs demonstrate the highest levels of agreement on entailment claims, followed by numeric errors and opposite statements. In contrast, higher rates of disagreement are observed for other claim types, particularly for missing information as well as unrelated claims. The overall agreement rate for this subtask is 0.44, which is lower than the rate observed in subtask 1, indicating the increased complexity of the classification. Similarly, the Cohen's Kappa coefficient is 0.23, reflecting a fair but lower level of agreement compared to subtask 1. As mentioned previously, a third SME was included to account for the agreement rate and to adjudicate disagreements and aggregate labels for these results, therefore ensure a higher label quality.

Label	Count	Percentage
entail	1762	38.74%
relunvef	926	20.34%
negat	701	15.40%
misinter	493	10.83%
entierr	256	5.63%
unrelunvef	185	4.07%
numerr	132	2.90%
missinfo	95	2.09%

Table 4: Claim distribution for the full Subtask 2 dataset.

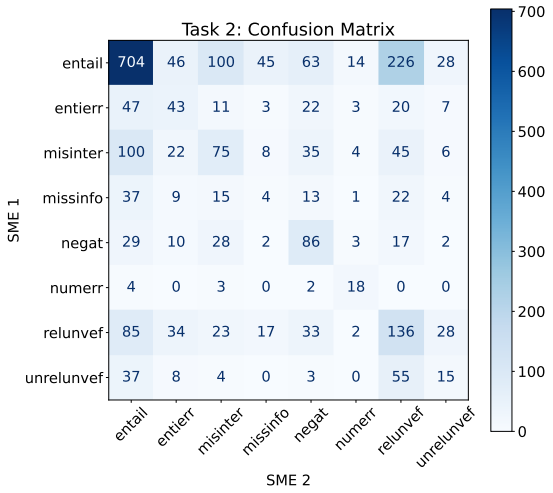

Figure 3: Confusion matrix comparing the predictions of two independent SMEs for Subtask 2. This figure is based solely on a more challenging subset of the data, comprising approximately 50% of the entire dataset, for which labels from the second SME are available.

	Precision	Recall	F1	Support
entail	0.57	0.67	0.62	1043
entierr	0.28	0.25	0.26	172
misinter	0.25	0.29	0.27	259
missinfo	0.04	0.05	0.04	79
negat	0.49	0.33	0.40	257
numerr	0.67	0.40	0.50	45
relunvef	0.38	0.26	0.31	521
unrelunvef	0.12	0.17	0.14	90

Table 5: Classification report comparing the predictions of two independent SMEs for Subtask 2. This report is based solely on a more challenging subset of the data, comprising approximately 50% of the entire dataset, for which labels from the second SME are available.

6 Competition Setup

6.1 Task

The Hallucination Detection for Scientific Content (SciHal) task challenges participants to identify hallucinated claims within answers generated by GenAI-powered research assistants in response to research-oriented questions. Formulated as a multi-label classification problem, each instance includes a question, a generated answer, an extracted claim, and a set of reference abstracts. The objective is to classify each claim based on its alignment with the reference abstracts, using a predefined set of hallucination types.

The task consists of two subtasks. **Subtask 1:**

Coarse-grained Hallucination Detection requires classifying each claim into one of three categories: *Entailment*, *Contradiction*, or *Unverifiable*. **Subtask 2: Fine-grained Hallucination Detection** extends this framework by introducing a more detailed taxonomy, including the following labels: *Entailment*, *Unrelated and Unverifiable*, *Related but Verifiable*, *Misrepresentation*, *Missing Information*, *Numeric Error*, *Entity Error*, and *Opposite Meaning*.

6.2 Data

Table 6 lists data splits. Each data instance includes the following fields:

- `ID` – Unique identifier.
- `question` – The research-oriented question.
- `answer` – The answer generated by a GenAI-powered research assistant.
- `claim` – One or more sentences extracted from the generated answer.
- `reference` – One or more reference abstracts retrieved for grounding.
- `label` – The classification label (available only in training sets). Labels follow a three-class scheme for Sub-task 1 and an eight-class scheme for Subtask 2.
- `justification` – The reasoning provided by subject-matter experts (SMEs) for assigning the label (available only in training sets).

Dataset	# Claim
training	3592
validation	500
test	500

Table 6: The statistics of the training, validation, and test set.

6.3 Evaluaiton Metrics

The competition will use one of the default classification metrics on Kaggle - the weighted F1 score as the major evaluation metric. Weighted F1 calculates metrics for each label, and finds their average weighted by support (the number of true instances for each label). This alters 'macro' to account for label imbalance; it can result in an F-score that is not between precision and recall.

Team	Wt F1
ScaDS.AI x sebis (Schopf et al., 2025)	0.62
YupengCao (Cao et al., 2025)	0.60
Daria Galimzianova (Galimzianova et al., 2025)	0.59
A.M.P (Le and Thin, 2025)	0.58
Crivoi Carla (Carla and Uban, 2025)	0.56
Ioan-Cristian Cordos	0.47
sasha boriskin	0.46
Andreea Brandiburu	0.44
e0nia	0.43
JB	0.27

Table 7: Performance of participants on the test set on Subtask 1.

Team	Wt F1
ScaDS.AI x sebis (Schopf et al., 2025)	0.47
A.M.P (Le and Thin, 2025)	0.47
JB	0.47
YupengCao (Cao et al., 2025)	0.47
Crivoi Carla (Carla and Uban, 2025)	0.46

Table 8: Performance of participants on the test set on Subtask 2.

7 Result

Tables 7, 8, 9, and 10 list the results of participants on the validation and test sets. lists the results of participants on the test set.

5 papers got accepted at the Fifth Scholarly Document Processing workshop(Ghosal et al., 2025). In Schopf et al. (2025), the team framed hallucination detection as a Natural Language Inference (NLI) problem. Their approach leveraged fine-tuned transformer models—specifically ModernBERT and DeBERTa-v3-large, and combined them using a weighted ensemble. Their results demonstrate that fine-tuned NLI models can outperform prompting-based approaches. They also highlight the importance of training on data that closely resembles the target task.

Cao et al. (2025) proposed a hybrid hallucination detection system combining prompting strategies with internal state classification. They benchmarked LLMs using zero-shot and few-shot prompts with Chain-of-Thought reasoning, and found that instruction-tuned, larger models per-

Team	Wt F1
ScaDS.AI x sebis (Schopf et al., 2025)	0.60
YupengCao (Cao et al., 2025)	0.59
A.M.P (Le and Thin, 2025)	0.59
Daria Galimzianova (Galimzianova et al., 2025)	0.58
Crivoi Carla (Carla and Uban, 2025)	0.51
Andreea Brandiburu	0.46
Ioan-Cristian Cordos	0.46
sasha boriskin	0.45
e0nia	0.42
JB	0.25

Table 9: Performance of participants on the validation set on Subtask 1.

Team	Wt F1
YupengCao (Cao et al., 2025)	0.51
ScaDS.AI x sebis (Schopf et al., 2025)	0.50
JB	0.49
A.M.P (Le and Thin, 2025)	0.48
Crivoi Carla (Carla and Uban, 2025)	0.43

Table 10: Performance of participants on the validation set on Subtask 2.

formed best. To further enhance detection, they extracted LLM hidden states and trained a logistic regression classifier without fine-tuning the models. This approach achieved top leaderboard scores (0.59 on subtask 1, 0.51 on subtask 2), demonstrating the effectiveness of integrating prompt reasoning with representation learning.

Le and Thin (2025) proposed a hallucination detection system using prompt-engineered LLMs. They designed structured prompts with role definitions, label explanations, and few-shot examples, and introduced a two-step method that predicts fine-grained labels before mapping to coarse ones. This approach outperformed direct prediction, with their best model (gemini-2.5-flash) achieving weighted F1-scores of 0.56 (subtask 1) and 0.44 (subtask 2).

Galimzianova et al. (2025) approached coarse-grained hallucination detection as an NLI task. They found that simply fine-tuning NLI-pretrained encoders like DeBERTa-v3 on the task dataset outperformed more complex pipelines and prompting-based methods. The study reaffirms that, for small-scale, domain-specific scientific data, targeted encoder fine-tuning remains both effective and efficient.

Carla and Uban (2025) combined SciBERT with contrastive learning techniques to improve hallucination detection. They applied a dual-head architecture with classification and contrastive objectives, using both Triplet and InfoNCE losses alongside standard cross-entropy. Their method aimed to enhance semantic alignment between claims and references, especially when surface wording differs.

8 Discussion

The SciHal shared task attracted a wide range of approaches to hallucination detection in scientific content. The participating teams explored diverse techniques including prompt-based LLMs, fine-tuned encoders, hybrid fusion strategies, and internal state modeling. Top-performing systems consistently relied on fine-tuning transformer models. This outcome suggests that supervised adaptation remains effective in domains with limited training data and high factual precision requirements.

Annotation quality remains a key challenge. Despite expert annotators and adjudication, hallucination labeling involves subjectivity and is time-intensive — each claim required an average of 7 minutes to annotate. This underscores the need for more scalable and consistent annotation protocols.

Although fine-grained hallucination types are difficult to annotate, they are particularly valuable for real-world applications, as they reflect common failure modes observed in practical GenAI systems. These include, but are not limited to: non-synonymous term substitutions, suboptimal grounding, direct copying instead of summarization, over-generalization from a single source, tangential continuations, avoidance of direct answers, conceptual conflation, evidence overstatement. Capturing these phenomena offers critical insights into model behavior. However, such cases are relatively rare, making it challenging to collect sufficient labeled instances. This rarity, combined with the nuanced nature of these errors, also poses significant challenges for future work.

Another limitation lies in the data split strategy. The train/val/test sets were divided by claim rather than by question, resulting in all test questions being seen during training. However, the claims in train are very different than test: less than 1.5%

of test claims showed high similarity to training claims. Repeated exposure to identical questions and answer contexts may still favor memorization, particularly for fine-tuned models. Future iterations should ensure both question- and context-disjoint splits to better assess generalization.

In future work, we aim to expand hallucination type coverage, improve annotation consistency, and adopt stricter data partitioning to enable more robust benchmarking.

Acknowledgments

We thank the following individuals for their valuable contributions: Johanna Sergent, Joo Sic Choi, and Jaiganesh Subramanian for coordinating data annotation; Poonam Pandey, Akila Chandrasekhar, Veronique Moore, and Alamelu Mangai Krishnamurthy for reviewing the annotation guidelines and conducting trial annotations; Maya Oded, Alex Riemer for early discussions on the hallucination taxonomy; Jan Bij de Weg, Debarati Banerjee, and Ben Buckley for supporting the legal coordination of data release.

References

Yupeng Cao, Chun-Nam Yu, and K.P. Subbalakshmi. 2025. Detecting hallucinations in scientific claims by combining prompting strategies and internal state classification. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*.

Crivoi Carla and Ana Sabina Uban. 2025. Scibert meets contrastive learning: A solution for scientific hallucination detection. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*.

Daria Galimzianova, Aleksandr Boriskin, and Grigory Arshinov. 2025. From rag to reality: Coarse-grained hallucination detection via nli fine-tuning. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*.

Tirthankar Ghosal, Philipp Mayr, Anita de Waard, Aakanksha Naik, Amanpreet Singh, Dayne Freitage, Georg Rehmf, Sonja Schimmlerh, and Dan Li. 2025. Overview of the fifth workshop on scholarly document processing. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of trec 2024 biomedical generative retrieval (biogen) track. *arXiv preprint arXiv:2411.18069*.

Mengya Hu, Rui Xu, Deren Lei, Yaxi Li, Mingyu Wang, Emily Ching, Eslam Kamal, and Alex Deng. 2024. Slm meets llm: Balancing latency, interpretability and consistency in hallucination detection. *arXiv preprint arXiv:2408.12748*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. *arXiv preprint arXiv:2311.05232*.

Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 525–534.

Khoa Nguyen-Anh Le and Dang Van Thin. 2025. A.m.p at scihal2025: Automated hallucination detection in scientific content via llms and prompt engineering. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*.

J Li, J Chen, R Ren, X Cheng, WX Zhao, JY Nie, and JR Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. arxiv, article. *arXiv preprint arXiv:2401.03205*.

Timothee Mickus, Elaine Zosa, Raúl Vázquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. Semeval-2024 task 6: Shroom, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM_Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. Association for Computational Linguistics.

Tim Schopf, Juraj Vladika, Michael Färber, and Florian Matthes. 2025. Natural language inference finetuning for scientific hallucination detection. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.