From RAG to Reality: Coarse-Grained Hallucination Detection via NLI Fine-Tuning

Daria Galimzianova, Aleksandr Boriskin, Grigory Arshinov
MTS AI

Abstract

We present our submission to SciHal Subtask 1: coarse-grained hallucination detection for scientific question answering. We frame hallucination detection as an NLI-style three-way classification (entailment, contradiction, unverifiable) and show that simple fine-tuning of NLI-adapted encoder models on task data outperforms more elaborate feature-based pipelines and large language model prompting. In particular, DeBERTa-V3-large, a model pretrained on five diverse NLI corpora, achieves one of the highest weighted F1 scores on the public leaderboard. We additionally explore a pipeline combining joint claim–reference embeddings and NLI softmax probabilities fed into a classifier, but find its performance consistently below direct encoder fine-tuning. Our findings demonstrate that, for reference-grounded hallucination detection, targeted encoder fine-tuning remains a competitive approach.

1 Introduction

Generative AI assistants are increasingly utilized to produce reference-based answers in scientific and research contexts, particularly via retrieval-augmented generation (RAG) systems that combine large language models with external knowledge sources. While RAG can greatly improve factual coverage, it also introduces a critical problem: *hallucinations*, wherein the model generates claims that are unsupported or directly contradicted by the cited references. Detecting such hallucinations is essential for trustworthy scientific communication, yet remains a major challenge for evaluation pipelines.

The **SciHal** shared task on **Hallucination Detection for Scientific Content** (Li et al., 2025), organized at Workshop on Scholarly Document Processing at ACL 2025, formalizes this problem as a multi-label classification task. Given a research question, a model-generated summary, and an extracted claim, participants must determine whether each claim is *entailment*, *contradiction*, or *unverifiable* with respect to the provided reference abstracts (Subtask 1: Coarse-Grained Detection).

In this paper, we report our approach that was ranked fourth and share our experiments for SciHal Subtask 1. We explore three families of approaches:

1. **Cross-Encoder Fine-Tuning**: We adapt NLI-pretrained encoders (most notably DeBERTa-v3) directly on the Subtask 1 training data, achieving a competitive score of weighted F1 (0.58).

2. **Feature-Based Classification Pipelines**: We experiment with semantic similarity features and NLI probability scores to train a classifier for label prediction. While more computationally intensive, these pipelines underperform relative to specialized encoder fine-tuning.

3. **LLM Prompting**: we deploy large language models like Qwen in a few-shot setting, which do not yield any promising results for the claim classification task.

Our analysis shows that among the methods listed above, *straightforward fine-tuning of an NLI-adapted encoder* yields the best performance on the task test dataset. We conclude that, for coarse-grained hallucination detection, simpler encoder-only architectures might be an efficient choice.

2 Related work

Automatic verification of factual consistency has attracted intense attention in recent years, reflecting a surge of methods and datasets devoted to achieving this goal (Li et al., 2022). Within the scientific domain, this research tradition is grounded in claim-verification datasets such as SciFact (Wadden et al., 2020) and follow-up shared tasks like SCIVER

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 353–359

(Wadden and Lo, 2021). The NAACL SCIVER scaled verification to a 5 M-abstract corpus, systems had to retrieve evidence and assign support or refute labels. All top-3 teams combined sparse retrieval with domain specific BERT-family encoders like SciBert (Beltagy et al., 2019a), BioBert (Lee et al., 2020) and Roberta (Liu et al., 2019). Recent developments (Mor-Lan and Levi, 2024) (Sankararaman et al., 2024) show that NLI cross-encoders demonstrate the ability to discern factual claims from non-factual ones given evidence. NLI is the task of determining whether a "hypothesis" text can be inferred (entailment), contradicted (contradiction), or is undetermined (neutral) from a "premise" text. NLI cross-encoders (Mor-Lan and Levi, 2024) are a natural fit for scientific claim verification because their label space: entailment, contradiction, and neutral is isomorphic to the support, refute, unverifiable taxonomy used in SciFact, SCIVER and SciHal Subtask 1, so no label remapping is required.

3 Data

The dataset is a claim-level annotated benchmark designed to measure factual correctness and hallucination detection quality in scientific RAG systems. Data was originally sourced from logs of a scientific research assistant tool. 50,000 samples were collected over one week. Organizers used an LLM to classify user questions by domain, ensuring completeness and correctness, and filtered out non-English texts.

Given the feasibility of employing subject matter experts (SMEs) for annotation, the dataset authors included texts from multiple domains: Engineering, Environmental Science, Medicine, Agricultural and Biological Sciences, and Computer Science. All questions were rewritten using an LLM, after which human annotators removed confidential or commercial information. This process yielded a refined dataset of 500 samples. Organizers then used a RAG scientific research assistant to retrieve 20 relevant article abstracts for each question. Finally, they generated answers and extracted claims with corresponding references.

To balance class distribution, the authors prompted an LLM to synthetically falsify claims. They modified 75% of samples by corrupting claims according to predefined fallacy types (for Sub-task 2), while maintaining class balance. The resulting dataset was manually verified and annotated by SMEs. This methodology ensured only 25% of samples were marked as entailment, with other classes each representing less than 10% of samples. Synthetic corruption also reduced manual annotation costs.

In the first annotation round, one SME validated samples within their domain, establishing baseline human annotations. Samples where both the LLM and SME agreed formed Batches 1 and 2. The remaining samples were annotated by a second SME. To achieve consensus, a third SME made final label decisions after reviewing justifications from prior annotators. This process improved the quality and challenge level of Batch 3 and the test set. The resulting Batch 2 (that also includes Batch 1 data), Batch 3 and test set contain 2092, 1500 and 1000 samples accordingly. Each record contains: the original question, the AI-generated answer, one or more claims (extracted from the answer), one or more references (article abstracts from the RAG tool), a label (for training sets only; three-class scheme: *entail, unver, contra*) and justification (SME reasoning for the label; training sets only).

4 Experiments

We approach this task as a pure classification problem, namely we view it as a NLI task. All three NLI labels are identical in their sense to the labels provided in the training data of this task.

Given limited time and resources, we train three groups of models. The choice of BERT-like models fine-tuned for the NLI task is intuitive, since the labels we need to predict are directly used in NLI. To test a more complicated pipeline, we train a classification model on features extracted from the data: cosine similarity scores for the embedded claims and references and probabilities for NLI classes. The third group is LLM-based: we prompt Qwen3-8B [1] to predict the labels with two examples for each class taken from the training set. The resulting scores can be found in Table 1.

4.1 BERT-based Models

We fine-tune four BERT-based models on the training data provided by the organizers. To adapt the standard NLI annotation scheme, we map the traditional NLI *neutral* label to the *unverifiable* label used in this task. Each model has been previously fine-tuned on multiple generic NLI datasets.

[1] https://huggingface.co/Qwen/Qwen3-8B

Model	F1 Weighted
BERT-based	
SciBERT-NLI	0.35
ModernBERT-base	0.56
ModernBERT-large	0.57
DeBERTa-NLI	**0.58**
Classifiers	
SciBERT + DeBERTa-NLI	0.34
SciBERT-FT + DeBERTa-NLI	0.51
Few-shot LLMs	
Qwen3 8B	0.45

Table 1: F1 weighted scores reported on the public leaderboard for various approaches.

- **SciBERT-NLI** (Beltagy et al., 2019b)[2]: Chosen for its domain-specific vocabulary and prior adaptation to scientific NLI.
- **ModernBERT-NLI** (Sileo, 2024): Built on the ModernBERT architecture, supports long contexts (up to 8 192 tokens) and trained on a blend of NLI corpora.
- **DeBERTa-V3-large-NLI** (Laurer et al., 2023)[3]: Pre-fine-tuned on over 800 k hypothesis–premise pairs, yielding the strongest performance on the leaderboard.

4.2 Classification Models

To capture richer signals from NLI models, we construct a secondary pipeline that treats NLI outputs and embedding similarities as features for a downstream classifier:

1. **Embedding Similarity.** Embed each claim and each reference abstract with SciBERT (chosen for its scientific domain fit) and compute their cosine similarity.
2. **NLI Probabilities.** Run DeBERTa-V3-large-NLI on each (reference, claim) pair and collect the softmax probabilities for entailment, contradiction, and unverifiable.
3. **Feature Concatenation.** Concatenate the cosine similarity scores and NLI probabilities into a single feature vector for each claim.
4. **CatBoost Classification.** Train a CatBoost classifier (Dorogush et al., 2018) on these feature vectors to predict the three coarse-grained labels.

We also experiment with fine-tuning SciBERT on the claim–reference classification task prior to feature extraction. In this variant, we:

- Extract `[CLS]` embeddings from the fine-tuned SciBERT model for every claim and reference.
- Recompute cosine similarities using these task-adapted embeddings.
- Apply the same CatBoost pipeline with DeBERTa-NLI probabilities.

This enhanced feature pipeline improves over the vanilla version, but still underperforms compared to direct fine-tuning of DeBERTa for Subtask 1.

4.3 Embedding Visualization

To assess whether our joint (reference, claim) embeddings capture class-discriminative structure, we computed `[CLS]` vectors from our fine-tuned SciBERT-NLI model for each claim–reference pair and aggregated them by mean. We then measured the silhouette score on these high-dimensional embeddings (silhouette = 0.0885), indicating poor cluster separation. A 2D t-SNE projection (Figure 1) further confirms that the three classes (entailment, contradiction, unverifiable) do not form well-separated clusters. This suggests that even with task-specific fine-tuning the learned embedding space may not suffice for clear, unsupervised clustering of hallucination types.

4.4 LLMs

For the LLM setting, we apply the Qwen3-8B (Yang et al., 2025) model with few-shot demonstrations per class (entail, contra, unver) drawn from the training set. Each inference prompt consists of a question, an answer, a claim and 2 references separated by `[SEP]` tokens. We evaluate Qwen3-8B only with non-thinking mode. The prompt template can be found in Figure 2.

5 Error analysis

We conduct a qualitative error analysis on a development set comprising 360 samples (10% of the training data). Table 2 presents the classification performance of the DeBERTa-V3-large model

[2] https://huggingface.co/gsarti/scibert-nli
[3] https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

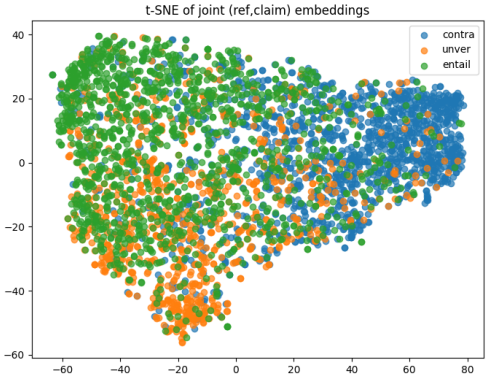

Figure 1: t-SNE projection of joint (reference, claim) embeddings from the fine-tuned SciBERT-NLI model, colored by true label (entailment, contradiction, unverifiable). The lack of well-separated clusters corroborates the low silhouette score.

fine-tuned for one epoch. The majority of errors involved confusion between the *entailment* and *unverifiable* classes, with *unverifiable* often incorrectly classified as *entailment* (32.58% of all misclassifications). This indicates that the model tends to assume textual support even in uncertain scenarios.

Class	**Precision**	**Recall**	**F1**	**Support**
Unver	0.73	0.60	0.65	89
Contra	0.84	0.82	0.83	137
Entail	0.70	0.79	0.74	134
Accuracy			0.76	360
Macro Avg	0.75	0.74	0.74	360
Weighted Avg	0.76	0.76	0.75	360

Table 2: Fine-tuned DeBERTa-V3-large model (scores on the development set).

Output Class \ Target Class	contra	entail	unver
contra	0 0%	17 37%	7 35%
entail	15 68%	0 0%	13 65%
unver	7 32%	29 63%	0 0%

Table 3: Misclassification matrix of DeBERTa-V3-large (counts).

Table 3 illustrates specific error patterns, emphasizing that misclassifications predominantly involve confusion between *entailment* and *unverifiable*. This suggests that improving the model's discrimination between supported and uncertain statements could substantially enhance performance.

We also evaluated the CatBoost-based pipeline with similarity and NLI features, which demonstrated notably lower performance (weighted F1: 0.57 on the development set), primarily due to increased confusion across classes, particularly between *contradiction* and *entailment*.

Additionally, we assessed the Qwen model, which exhibited a significantly higher error rate (43.4%) on a similar development set. The Qwen model predominantly confused *contradiction* with *entailment* and vice versa, highlighting fundamental issues in distinguishing these classes effectively. This suggests that the Qwen model requires further adaptation or training enhancements to reliably detect textual support and contradiction in scientific contexts.

Some classification examples can be found in Table 4.

6 Results

The resulting weighted F1 scores on the public leaderboard are shown in Table 1. The most efficient and highest-performing approach is simply fine-tuning NLI-adapted encoder models on the task's training data. Interestingly, DeBERTa-v3 (released in 2022 and fine-tuned on five diverse NLI datasets) outperforms the newer ModernBERT, despite the latter having seen a much larger set of NLI pairs. Our fine-tuned DeBERTa model therefore ranks fourth on the public leaderboard for Subtask 1.

By contrast, more elaborate pipelines that extract features via embeddings and NLI probability outputs incur substantial computational overhead and still underperform compared to direct cross-encoder fine-tuning.

Similarly, in-context learning with LLM that we evaluated, while a popular choice, is markedly more expensive to run and achieves lower F1 scores than the smaller, specialized encoder models.

These results resurface the importance of task-specific transfer learning and highlight the role of training data quality over multi-domain generalization abilities of a model.

```
You are a scientific claim validator for the SciHal task.
Given (1) Question, (2) Claim, (3) References (abstracts)
decide which label applies following the decision tree:
• If the paragraph is not well-formed filler → output "unver".
• Else, if claim is ENTIRELY supported by 1 abstract AND not contradicted by any other → "entail".
• Else, if claim is DIRECTLY contradicted (different number/entity/relation) → "contra".
• Else → "unver".
Return ONLY one of: entail | contra | unver

FEW_SHOT_BLOCK

Question: {row.query.strip()}
Answer snippet: {answer_snippet}
Claim: {row.claim.strip()}
References: {refs}
Label:
```

Figure 2: Prompt template used for Qwen3-8B.

7 Conclusion

We have addressed the challenging problem of coarse-grained hallucination detection in scientific question answering by framing it as a three-way NLI task (entailment, contradiction, unverifiable) on retrieved reference abstracts. We evaluated a range of methods—from simple encoder fine-tuning (SciBERT, ModernBERT, and DeBERTa-V3) to more elaborate feature-based pipelines combining joint claim–reference embeddings and NLI softmax scores, as well as few-shot prompting of large language models. Our experiments demonstrate that direct fine-tuning of an NLI-adapted cross-encoder, and in particular DeBERTa-V3-large, offers competitive accuracy in solving the hallucination detection task, achieving fourth place on the public leaderboard for SciHal Subtask 1. The modest performance of unsupervised embedding clustering and the underwhelming results of more complex pipelines underscore the inherent difficulty of reliably detecting scientific hallucinations without explicit supervision. Our findings reaffirm that, despite the task's complexity, targeted cross-encoder fine-tuning remains an effective strategy for reference-grounded hallucination detection.

8 Limitations

While our study demonstrates the strong performance of NLI–adapted cross-encoder fine-tuning for coarse-grained hallucination detection, several limitations remain:

- **Model diversity.** We evaluated a relatively small set of encoder models (SciBERT, ModernBERT, DeBERTa-V3). Exploring additional architectures, especially lighter or multilingual encoders, may yield further gains.
- **Data scale.** Although our training splits are carefully annotated and of high quality, the overall dataset remains modest in size. Larger or more varied annotated corpora could improve robustness and generalization.
- **LLM fine-tuning.** We only tested few-shot prompting of large language models. With task-specific fine-tuning and structured-output prompts (e.g. chain-of-thought templates), LLMs may ultimately surpass encoder-only approaches, at the expense of greater computational cost.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019a. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019b. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, pages 1–33.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Dan Li, Bogdan Palfi, and Colin Kehang Zhang. 2025. Hallucination detection for scientific content. https://kaggle.com/competitions/hallucination-detection-scientific-content-2025. Kaggle.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Guy Mor-Lan and Effi Levi. 2024. Exploring factual entailment with nli: A news media study. *arXiv preprint arXiv:2406.16842*.

Hithesh Sankararaman, Mohammed Nasheed Yasin, Tanner Sorensen, Alessandro Di Bari, and Andreas Stolcke. 2024. Provenance: A light-weight fact-checker for retrieval augmented llm generation output. *arXiv preprint arXiv:2411.01022*.

Damien Sileo. 2024. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

David Wadden and Kyle Lo. 2021. Overview and insights from the sciver shared task on scientific claim verification. *arXiv preprint arXiv:2107.08188*.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Claim	Justification	DeBERTa	CatBoost	Qwen	True Label
Common Antibiotics Detected: Studies have identified various antibiotics in water sources, including tetracyclines, sulfonamides, and quinolones, suggesting that all water sources are likely contaminated with these substances at harmful levels [2, 3, 4].	The claim generalized that all water sources are likely contaminated with various antibiotics; however, the cited references specifically mentioned contamination only in engineered aquatic environments.	entail	entail	entail	contra
Determination of Optimal Conditions: Use the model to determine the optimal conditions that maximize the desired responses. For example, optimal conditions might include specific temperature, time, and solvent concentration that yield the highest antioxidant activity [1, 2, 3, 5, 6].	The claim and the experimental data should fit a second-order polynomial model with a high R, aligning with the methodologies and results described in the reference.	unver	entail	entail	entail
Heat Stress and Diet Composition: The Temperature-Humidity Index (THI) significantly impacts both water intake and DMI in dairy cows, with higher THI leading to increased water intake and decreased DMI [3]. Although this study focuses on cows, similar effects can be expected in goats, suggesting that environmental conditions and diet composition are crucial factors in managing water and dry matter intake.	Both the claim and reference address how heat stress affects diet composition of livestock. The reference correlates it with lactating dairy cows. However, the claim implies that goats can have similar affects which the reference did not mention.	entail	contra	unver	unver

Table 4: Classification examples from the development dataset.