RaggedyFive at SemEval-2025 Task 3: Hallucination Span Detection Using Unverifiable Answer Detection

Wessel Heerema, Collin Krooneman, Simon van Loon, Jelmer Top, Maurice Voors
University of Groningen
(w.heerema, c.krooneman, s.p.c.a.van.loon, j.top, m.voors)@student.rug.nl

Abstract

This paper describes our submission to the SemEval-2025 Task 3: Mu-SHROOM, a shared task focused on hallucination span detection in the outputs of large language models (LLMs). The goal of the task is to identify spans of text that, despite being grammatically sound, are not supported by external sources. As a baseline, we employed random and zero-probability classifiers to gauge the difficulty of the task. Our main system combines a Retrieval-Augmented Generation (RAG) module with a Natural Language Inference (NLI) model to detect hallucinated spans. The RAG module retrieves information from Wikipedia and generates a premise, which is then compared to the LLM output using a multilingual NLI model in a sliding window approach. Our final system achieved competitive results, demonstrating the effectiveness of integrating RAG with NLI for fine-grained hallucination detection.

1 Introduction

Large language models (LLMs) are specialized in generating human-like text in various styles, which lends them to many practical applications. However, even the most sophisticated models can produce hallucinations, making users question their reliability and putting the adoption of machine learning pipelines in jeopardy (Rykov et al., 2024). Hallucination refers to the generation of texts or responses that exhibit grammatical correctness, fluency, and authenticity, but deviate from the provided source inputs or do not align with factual accuracy (Ji et al., 2023; Ye et al., 2023). This is a phenomenon that established evaluation metrics struggle to detect (Bahad et al., 2024); as a result, it has now become imperative to develop systems that can assess the factual consistency of a claim with respect to context (Zha et al., 2023).

The SemEval shared task Mu-SHROOM (Vázquez et al., 2025) provides an opportunity to develop solutions to the problem of hallucinations in LLMs. The objective of the task is to classify spans, which are continuous segments of text within LLM outputs, as hallucinations. For instance, in the generated sentence "Marie Curie won three Nobel Prizes for her work in physics and chemistry," the span "won three Nobel Prizes" would be labeled as a hallucination, since she actually won two. To this end, we only consider spans hallucinated when the LLM output contradicts the relevant retrievable information. The detection of hallucination spans allows for a more fine-grained understanding of where hallucinations occur in LLMs, as well as giving an indication of the severity of hallucinations in LLMs. This is something that a binary classification is not able to provide, given its simplicity.

In this paper, we present a linear composite system that employs a Retrieval-Augmented Generation (RAG) question-and-answer system to generate an answer to the question in each prompt; combines the question-answer pair into a unified premise with a generative LLM; and compares this premise with the subject model output using an off-the-shelf Natural Language Inference (NLI) model. We compare the performance of this system with several baselines, and discuss its strengths and weaknesses.

2 Background

The Shared-task on Hallucinations and Related Observable Overgeneration Mistakes (SHROOM) has been put forth by Mickus et al. (2024) to address the issue of hallucinations in LLMs. The main objective of SHROOM is the development of systems that detect hallucinations in the generated output of LLMs. In the shared task, participants must detect grammatically sound outputs that nonetheless contain incorrect or unsupported semantic information compared to a source input. This task can be done with or without access to the model that produced

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1473–1478

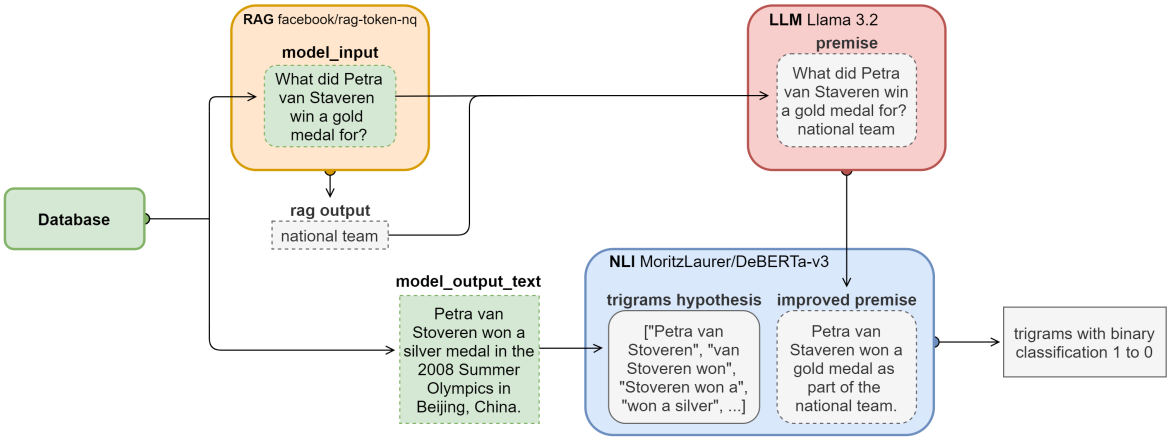

Figure 1: Overview of the improved and final system pipeline.

the output: respectively, these are the model-aware and model-agnostic versions of the task (Bahad et al., 2024; Mickus et al., 2024).

In this previous shared task, NLI-based approaches achieved the best performance. (Maksimov et al., 2024; Obiso et al., 2024). The objective of NLI systems is to determine the truth value of a hypothesis, given a premise. As an example, the premise *"the pedestrian walks on the zebra crossing"* and the hypothesis *"the pedestrian must yield"* produces a contradiction and is judged false; the same premise with the hypothesis *"the pedestrian is wearing a green shirt"* results in a neutral judgment, though this can also be rendered as false.

The remarkable similarity between NLI and the SHROOM task lent itself to several submissions utilizing models that were (pre-)trained on NLI data. The DeepPavlov team of Maksimov et al. (2024) opted to directly train RoBERTa and similar models as well as a Text-to-Text Transfer Transformer on NLI data. On the other hand, the HaRMoNEE team of Obiso et al. (2024) selected a pre-trained RoBERTa model and fine-tuned it using data from SHROOM. The models produced by these teams achieved accuracies of 0.80 and higher, with HaRMoNEE's approach being the best model in the model-aware version of the task.

Systems that operate on a similar objective of NLI, such as those for information retrieval, paraphrasing, fact verification and textual similarity, can be unified under a single model for information alignment (Zha et al., 2023). RAG frameworks have shown promising results in regard to hallucinations. These systems combine generative models with retrieval mechanisms. This hybrid method not only improves the factual accuracy of the generated text, but also helps mitigate the risk of hallucinations by grounding the output in verifiable data (Lewis et al., 2020).

The system used in this paper will use the insights from these works to combine NLI with RAG to create a pipeline for hallucination detection on text spans. In this way, we hope to build upon the best-performing systems from the SHROOM task and test their resilience against a different method of hallucination detection.

3 Data

In the SemEval 2025 Task-3 Mu-SHROOM, the task is to detect hallucination spans in the outputs of instruction-tuned LLMs in a multilingual context. To this end, the data is provided in 24 different languages, with each output being produced by a variety of open-source LLMs. The LLM output is provided in the format of a human question, a generated answer and, in the case of the validation set, labels for hallucinated spans that use string positions. The last of these is divided into soft labels, which indicate the probability of hallucinations, and hard labels, which assert spans as hallucinated if the probability exceeded 50%.

Val	Test	Total
50	154	1013

Table 1: The English data distribution for the shared task.

We chose to focus on the English language for this study, which was split at around 75%, 5% and

15% for training, validation and testing data respectively. The database and dataset distribution as utilized in our study can be seen in Table 1. While we confirm that an unlabeled training set was available, we did not make use of this set in our system.

4 Method

Our approach can be distilled into three distinct steps. The pipeline for this system is shown in Figure 1. First, the RAG model retrieves the relevant information from a Wikipedia vector index based on the prompt question, and generates an appropriate answer to form a premise. Afterwards, our model tokenizes the hallucinated answer using the Treebank tokenizer from the NLTK library (Bird and Loper, 2004). Finally, the hallucinated answer is fed in token trigrams to the NLI model as hypotheses, with the trigrams being fed in a sliding-window fashion.

Besides our main system, we also created a baseline classifier that uses random probabilities and all-zero probabilities. The random-probability baseline classifier is the default and assigns completely random probabilities to each span, making each output unique and not reproducible. The alternative approach assigns a 0.0 value for all probabilities. These baselines are meant to gauge the effectiveness of our systems in the absence of external baseline metrics during development.

4.1 Retrieval-Augmented Generation

We employ the RAG system as designed by Lewis et al. (2020) and use the default wiki_dpr vector (Karpukhin et al., 2020) as its dataset. Due to computational constraints, we did not include a document screening stage to filter irrelevant or low-quality factual documents; this represents realistic limitations in low-resource settings.

The handling of the RAG output can occur in two ways. The more basic implementation concatenates the input question and the RAG answer to form the premise. In our main system, we employ Llama 3.2 (Dubey et al., 2024) running under the Ollama API to rewrite the concatenated premise into a natural-language answer to the question. The prompt used for this component, as well as the manner in which a question and answer pair is formulated in the prompt, can be found in Appendix A. We tested both approaches in this study.

4.2 Natural Language Inference

We employ an off-the-shelf multilingual DeBERTa NLI model that is fine-tuned on three datasets (Laurer et al., 2022; He et al., 2021, 2023), comprising 885 to 242 NLI hypothesis-premise pairs. We use this model as it was provided, without any additional fine-tuning on the provided hallucination detection dataset or other similar datasets.

The NLI model evaluates whether the claims in the generated response logically follow from the RAG premise, given in regression probabilities of entailment, neutrality and contradiction. If the contradiction probability of a trigram is at least 0.1 and it is larger than the entailment probability, then it scores that trigram as 1; otherwise, it is left as 0. The soft label probability for each token is the average classification of every trigram that the token occurs in. While this averaging method provides an intuitive and lightweight way to generate soft labels, we acknowledge that this differs from conventional hallucination detection practices.

4.3 Evaluation

To evaluate the performance of our system, Vázquez et al. (2025) have specified the use of an Intersection over Union (IoU) score and Spearman's correlation coefficient (ρ). The IoU is calculated on the index sets of hallucination spans between the gold reference and the predictions per hypothesis. If the calculated hallucination probability score of a span is greater than 0.5, the evaluation program classifies the span as a hallucination. The span is then converted to a set of indices. For instance, the soft label {"start": 32, "prob": 0.667, "end": 34} is transformed into the set {32, 33, 34}. The indices for all spans classified as hallucinations is combined into a single set. These sets are compared between the gold reference and the predictions of the models. To calculate ρ between the gold reference and the predictions, the evaluator program compares probability vectors for all of the spans.

5 Results

Our main system ranked 32nd on the English-language leaderboard, measured by the IoU score on the test set. We applied our baseline models to the validation set only; the baseline scores reported for the test set are provided by Vázquez et al. (2025). The full results are shown in Table 2.

The all-zero baseline resulted in an IoU score of

Validation	**IoU**	**Cor**
Baseline (all-zero)	0.040	0.000
Baseline (random)	0.187	0.179
System	0.198	0.171
System (+ Llama)	**0.240**	**0.201**
Test	**IoU**	**Cor**
Baseline (neural)	0.031	0.119
Baseline (mark none)	0.033	0.000
System	0.275	0.261
System (+ Llama)	0.315	**0.304**
Baseline (mark all)	**0.349**	0.000

Table 2: The results from our system on the validation and test sets, as compared to the available baseline systems. The scores with systems in italics were gathered from the leaderboard.

0.040 and a correlation score of 0.000, indicating that it fails to provide meaningful hallucination span predictions. The random baseline performed slightly better, achieving an IoU score of 0.187 and a correlation of 0.179, demonstrating that a completely random assignment can capture some degree of variation in hallucination spans, though it remains unreliable.

Our main model on the validation set achieved an IoU score of 0.198 and a correlation score of 0.171, showing a slight improvement over the baseline models. In the test set, our model demonstrated a larger increase in performance, with an IoU score of 0.275 and a correlation of 0.261. This suggests that our concatenation-based system yields an improvement in identifying hallucination spans beyond what is captured by baseline approaches; however, if this were the case, the effect size is negligible.

The improved system with Llama premise rewriting demonstrated the most visible gains in performance. On the validation set, the improved system achieved an IoU of 0.240 and a correlation of 0.201. The results for the improved system on the test set yielded our highest scores overall, with an IoU of 0.315 and a correlation of 0.304. These results indicate that the addition of a premise-rewriting step refines the hallucination detection process and leads to a more robust identification of hallucination spans.

6 Discussion

The results indicate that our proposed system provides a noticeable improvement over our baseline models. Using retrieval-augmented generation, the model ensures that the responses generated are grounded in relevant contextual information. Furthermore, NLI-based evaluation at the trigram level enables a more granular detection of hallucination spans, which is not possible with binary classification approaches.

A key strength of our approach lies in the introduction of a premise rewriting step, which improves alignment between generated text and factual sources before the NLI step. The empirical results show that this method enhances the detection performance of the hallucination range. However, this did not improve the detection of subtle hallucinations or the handling of paraphrased incorrect information.

Our core approach has a few systematic flaws worth addressing. For instance, we used an off-the-shelf NLI model without additional task-specific fine-tuning. While this allowed for rapid experimentation, it may have resulted in worse overall performance. Fine-tuning or adapting the model on the hallucination detection dataset itself could have improved performance by aligning the model more closely with task-specific patterns. A similar problem exists for our RAG component, whereby an unscreened set of documents may have led to lower-quality training data for the RAG model. We expect a custom, curated set of documents to improve the overall efficacy of the model.

In addition, our system is non-standard in ways that could affect performance. Many systems either use external factual documentation to explicitly verify claims or assess internal output consistency across different runs. In contrast, our method does not rely on external factual verification beyond the initial RAG retrieval, nor does it compare outputs across runs. Furthermore, the detection is entirely localized, whereby the contradiction entailment is combined with the span searching. This may also have contributed to a lower leaderboard rank.

We manually checked differences in span hallucination assessment for the validation set between the gold reference and our best model, in order to better understand the performance of our model. In particular, the span annotations in the gold standard are different from the spans that we created using the Treebank tokenizer. For instance, the first identified span in the sentence "The Elysiphale order contains 5 genera." is the word 'the' for our own system, and 'Elysiphale' for the gold reference. The last identified span is '.' and 'genera.'

respectively. In this example, the identified span 'genera' is assigned a probability of 1.0 by our own model, and the span 'genera.', including the period, is assigned 0.2 in the gold standard. If our NLI model classifies a trigram as a hallucination, that classification is extended to all of its constituents. This makes it prone to false positives, especially at string boundaries. We highlight an example of this behavior in Table 3.

Tokens	Predicted prob.	Gold prob.
The	0.0	0.0
Elysiphale	0.0	0.2
order	0.3333	0.0
contains	0.6666	0.0
5	1.0	1.0
genera	1.0	0.2
.	1.0	

Table 3: A comparison of the predicted and gold-standard soft labels for the sentence "The Elysiphale order contains 5 genera." The gold standard counts 'genera.' as a single token.

Finally, in the returned answers for the validation set, there were several questions that the RAG could not parse meaningfully. For instance, a query for the debut of Chance the Rapper returned his birth date, whereas a query for four elements in Zhejiang cuisine returned a single element. As a result, this augments only a part of the total output, instead of representing a fully augmented approach. Given that these answers were in a similar format to correct answers, we conclude that these are limitations of the RAG system itself and not the formulation of the questions. Future research could explore an alternative implementation that returns nearby answers in a JSON format, though the feasibility of this approach for vectors remains to be seen. Future work could also optimize the vector for an improvement in ease of use and deployment.

7 Conclusion

This paper presents a novel approach to hallucination span detection in machine-generated text through RAG and NLI. Our research is conducted within the framework of the Mu-SHROOM shared task, contributing to the broader effort of evaluating and improving hallucination detection techniques. Our results demonstrate that our proposed method outperforms baseline approaches and provides a more fine-grained understanding of hallucinations in LLM outputs. The introduction of a premise-rewriting step within the pipeline further enhances detection accuracy.

We recognize that our system has a variety of shortcomings that contributed to a lower score than most. In particular, future research could explore more selective labeling and a RAG-like system with an array of outputs. Nevertheless, we believe that our study contributes to the Mu-SHROOM shared task by providing information on hallucination span detection. In this way, we hope to advance research on the factual reliability of content generated by LLMs by mitigating the presentation of faulty information.

References

Sankalp Bahad, Yash Bhaskar, and Parameswari Krishnamurthy. 2024. NootNoot at SemEval-2024 task 6: Hallucinations and related observable overgeneration mistakes detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 964–968, Mexico City, Mexico. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Ivan Maksimov, Vasily Konovalov, and Andrei Glinskii. 2024. DeepPavlov at SemEval-2024 task 6: Detection of hallucinations and overgeneration mistakes with an ensemble of transformer-based models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 274–278, Mexico City, Mexico. Association for Computational Linguistics.

Timothee Mickus, Elaine Zosa, Raul Vazquez, 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, Mexico City, Mexico. Association for Computational Linguistics.

Timothy Obiso, Jingxuan Tu, and James Pustejovsky. 2024. HaRMoNEE at SemEval-2024 task 6: Tuning-based approaches to hallucination recognition. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1322–1331, Mexico City, Mexico. Association for Computational Linguistics.

Elisei Rykov, Yana Shishkina, Ksenia Petrushina, Ksenia Titova, Sergey Petrakov, and Alexander Panchenko. 2024. SmurfCat at SemEval-2024 task 6: Leveraging synthetic data for hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 869–880, Mexico City, Mexico. Association for Computational Linguistics.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona de Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 Task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *Preprint*, arXiv:2309.06794.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Llama prompt example

Your task is to rewrite a question and answer pair into a single, declarative sentence. Include all information from the original question, as well as information included in the answer. Both the question and the answer are provided below. Always assume the provided answer is correct. Do not include anything other than the resulting sentence in your response.

Question: What did Petra van Staveren win a gold medal for?
Answer: national team