Advacheck at SemEval-2025 Task 3: Combining NER and RAG to Spot Hallucinations in LLM Answers

Anastasia Voznyuk and **German Gritsai** and **Andrey Grabovoy**
Advacheck OÜ
{voznyuk, gritsai}@advacheck.com

Abstract

The Mu-SHROOM competition in the SemEval-2025 Task 3 aims to tackle the problem of detecting spans with hallucinations in texts, generated by Large Language Models (LLMs). Our developed system, submitted to this task, is a joint architecture that utilises Named Entity Recognition (NER), Retrieval-Augmented Generation (RAG) and LLMs to gather, compare and analyse information in the texts provided by organizers. We extract entities potentially capable of containing hallucinations with NER, aggregate relevant topics for them using RAG, then verify and provide a verdict on the extracted information using the LLMs. This approach allowed with a certain level of quality to find hallucinations not only in facts, but misspellings in names and titles, which was not always accepted by human annotators in ground truth markup. We also point out some inconsistencies within annotators spans, that perhaps affected scores of all participants.

1 Introduction

Modern advances in the field of text generation models provide artificial texts of a high quality that are hardly distinguishable from human-written texts at fluent reading. State-of-the-art instruction-tuned or reasoning Large Language Models (LLMs) are capable of generating an answer to any user query, but despite the attractiveness and conciseness of the answer, its appropriateness and accuracy often remained a controversial issue. LLMs are capable of retaining massive amounts of factual knowledge and use it to answer user queries, but they are still prone to generating hallucinations – statements that appear plausible but are factually incorrect or unverifiable (Maynez et al., 2020; Liu et al., 2023; Adlakha et al., 2024). Hallucinations pose a critical challenge in applications that demand high factual accuracy, such as medical or legal domains (Pal et al., 2023; Dahl et al., 2024). If a bit earlier we were concerned about detecting AI-generated content, nowadays with the increasing amount of the Internet being flooded with texts from LLMs, the focus on verification and validation of information is crucial (Gray, 2024; Gritsai et al., 2024). That brings us to the point where detecting and mitigating these hallucinations is essential for enhancing the reliability and trustworthiness of LLM outputs.

The SemEval 2025 Task 3 (Vázquez et al., 2025) poses a challenge in seeking hallucination within model answers on factual questions about famous people, locations or biological species. As most of these questions contain some sort of entity, we decided to employ Named Entity Recognition (NER) approach to determine for which entities there might be a relevant context. For determined entities we leverage Retrieval-Augmented Generation (RAG), as all questions from the dataset could be answered with context from Wikipedia. Relevant retrieved context together with model answer were given to LLMs that were tasked to evaluate the correctness of model's answer based on provided context and this were done twice obtain multiple opinions. After that, the final model had to edit the suggestions from LLM judges into initial answer. Through this method, we aim to detect and fix factual hallucinations from the text and thus contribute to ongoing efforts in hallucination detection and mitigation.

2 Related Work

Hallucination in LLMs refers to instances where models generate factually incorrect or unsubstantiated claims. Various methods have been proposed to detect and mitigate hallucinations, ranging from probabilistic confidence estimation (Manakul et al., 2023) to post-hoc verification using external knowledge sources. Self-consistency approaches have also been explored, where multiple model outputs

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1204–1210
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

are compared to detect inconsistencies. However, these approaches often struggle to identify hallucinations in complex, context-dependent settings. Hallucinations in LLMs can snowball when an initial false or misleading generation propagates through iterative interactions, reinforcing and expanding the error (Zhang et al., 2024). This happens when the model conditions future responses on its own prior outputs, amplifying inaccuracies over time. Therefore, it is important to detect hallucinations as soon as possible and abrupt model interaction with a context containing hallucinations.

2.1 Retrieval-Augmented Generation

Combining queries with additional information from sources can be valuable not only for preventing hallucinations in output, but also for recognising them. Retrieval-Augmented Generation (RAG) has emerged as a promising approach to improving factual consistency by integrating updated, relevant knowledge into the generation process (Lewis et al., 2020; Peng et al., 2023; Shuster et al., 2021). Studies have demonstrated that augmenting LLMs with structured or semi-structured data sources significantly reduces hallucination rates (Izacard et al., 2022). Nevertheless, even with RAG and other enhancements, LLMs still produce statements that are either unfounded or contradict the information provided in the retrieved references. It happens particularly when retrieval fails or retrieved documents contain inaccuracies (Gao et al., 2023). Authors of FAVA (Mishra et al., 2024) introduce a retrieval-augmented language model designed to detect and correct fine-grained hallucinations in generated text. They develop a benchmark comprising approximately one thousand human judgments across various domains, to assess hallucination detection performance. FAVA is trained using synthetic data specifically created to identify and rectify different types of hallucinations, significantly outperforming models such as ChatGPT and GPT-4 in both detection and factuality improvement tasks. In another paper authors introduce Dynamic Retrieval Augmentation (Su et al., 2024) based on hallucination Detection (DRAD) to address the posted issue. DRAD comprises two main components: Real-time Hallucination Detection (RHD), which identifies potential hallucinations during text generation without relying on external models, and Self-correction based on External Knowledge (SEK), which corrects detected inaccuracies by retrieving and incorporating relevant information from external sources. In our approach, we considered utilising RAGs in combination with LLMs as well.

2.2 Named Entity Recognition in Factual Consistency Evaluation

NER has been widely used in NLP for entity extraction and disambiguation, making it a useful tool for evaluating factual consistency in generated text (Wuehrl et al., 2023; Xie et al., 2023). Previous studies have shown that hallucinations often involve named entities being misrepresented or fabricated (Shen et al., 2023). By identifying and verifying named entities against reliable sources, NER-based methods can enhance hallucination detection.

3 Task Description

The objective of this task is to identify text spans that correspond to hallucinations in outputs, generated by LLMs. Provided data includes 14 languages and outputs from various publicly available LLMs. Participants receive an LLM-generated text in three formats:

- a raw character string
- a list of tokens
- a list of associated logits

Additionally, to determine spans more precisely, participants must assign a probability to each character or a span in the text, indicating the likelihood that it is part of a hallucination. There are two metrics: Intersection over Union (IoU) between predicted spans and ground-truth spans and Pearson correlation for evaluating how well the probability assigned by the participants' system correlates with the empirical probabilities observed by annotators. We have participated with our framework presenting approach for English, but it could be extended to other languages as well.

👱: Which mountain range is Speichersdorf located near?

🤖: Speichersdorf is located in the Black Forest mountain region of Germany.

Figure 2: Example of question and answer from the model with factual mistake (hallucination) coloured in red.

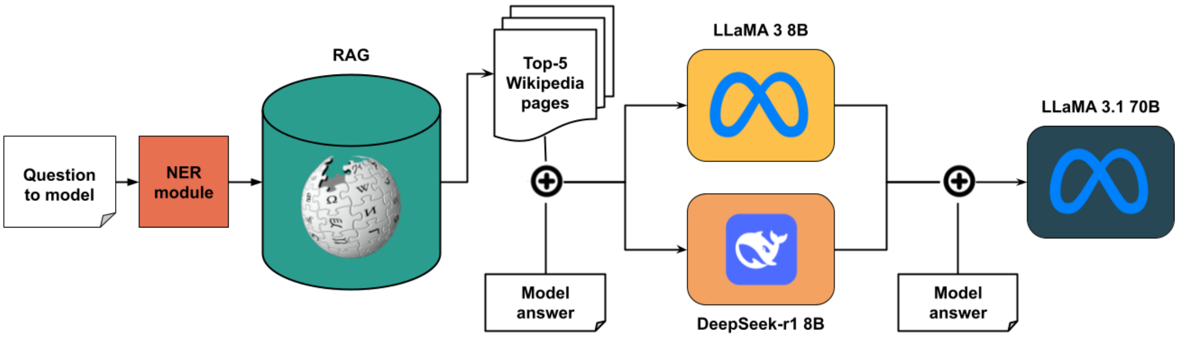

Figure 1: Our system pipeline.

4 Dataset

The dataset for the SHROOM 2025 challenge comprises a compilation of model-generated answers on factual questions. The goal is to find spans with hallucinations within model answer. Information for the data sample includes the following fields: (i) model input – the input question given to the generative language model; (ii) model id – the model used for generation; (iii) model output text – the text generated by a corresponding model; (iv) model output tokens – model output text, split to tokens; and finally (v) model output logits – logits for each generated token.

5 Proposed System

Our system employs a combination of Retrieval-Augmented Generation, Named Entity Recognition, and usage of LLMs to semi-automatically verify factual consistency and detect hallucinations in given answers. The workflow consists of the following key components:

1. NER module extracts entities from the question.
2. We find relevant to that entities Wikipedia pages through the RAG module.
3. Two small LLMs receive model answer and relevant Wikipedia context and are asked to factcheck it and output all factual mistakes.
4. Final judge LLM gets both outputs from small judges and must output the initial model answer with on-the-spot edits of factual mistakes.
5. Organisers' model answer and the answer of our system are compared to find differing spans.

It is illustrated in Figure 1.

6 Experimental Setup

We focused only on English subset, however, our system can be used for any language as long as LLMs support this language and there are knowledge sources in this language. To retrieve NER entities, we use DeepPavlov `ner_conll2003_bert` model (Savkin et al., 2024). We decided to extract sources from Wikipedia because after analysing the data provided, a significant amount of samples contain facts or references to various events. After retrieving the top-5 Wikipedia pages with the LangChain[1] loader and taking the most relevant one, we concatenate it with provided LLM answer and forward it to two LLM-judges: LLaMA3 8B (Grattafiori et al., 2024) and DeepSeek-R1 8B (DeepSeek-AI, 2025). Outputs from judges are concatenated with model answer and are given to the larger judge, LLaMA3.1 70B-Instruct. The latter judge should spot and edit the mistakes in the initial answer based on the smaller judges' opinions. Then, we check the difference between the original model answer and the edited answer. The spans that differed were stored and provided as the final answer. Additionally, we compared NER entities from the model answer and the question with cosine similarity, to find the inconsistencies and mistakes in the names.

[1] python.langchain.com/docs/introduction/

Model	**IoU**	ρ
Baseline (mark all)	0.348926	0
Advacheck (our)	0.44425	0.343241
Best Leaderboard	0.650899	0.629443

Table 1: Results from the leaderboard for English subset of the task.

7 Results and Discussion

Although our system's spans do not fully overlap with spans from annotators, with 0.44425 IoU score (see Table 1), we attribute some role of lower IoU to the subjectivity of the annotation. As it involves the work of human annotators, who are asked to provide their opinion about whether some span is hallucination, there is a discrepancy between what annotators marked as hallucinations and what we marked in our system, and also some level of inconsistency between different texts from test set. We noticed that the answers from the LLMs, provided by organisers, contained a lot of typos and mistakes in the name of people from the questions. For example, the question was about *Alberto Fouillioux* and the model begins its answer with the name *Albero Foulois*. We believe it to be an input-contradicting hallucination, whereas annotators did not mark it in this way, therefore almost all such cases negatively affected the score. At the same time, there are some examples where similar typos in the names were marked as hallucinations. In some cases, annotators aimed for precision by marking hallucinations at the word level, while in others, they annotated entire sentences. We provide more examples of inconsistencies in the Table 2. Also, once the model starts to hallucinate, it is more likely that it will continue hallucinating, as stated in Zhang et al. (2024).

Also, as our system employs LLMs, it is vulnerable to hallucinations from the LLM judges when they work with the provided context. The solution is to take an ensemble of weaker judges and a more capable final judge to eradicate possible appearing hallucinations.

8 Conclusion

We presented a system that combines NER and RAG modules and after that utilises LLMs to detect spans of hallucinated output and showed how it can be used to edit factual mistakes in model answers. It can be further modified to work with several knowledge sources or employ more LLM judges to obtain more opinions and decrease the risk of hallucinations of the judges.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, et al. 2024. The llama 3 herd of models.

Andrew Gray. 2024. Chatgpt "contamination": estimating the prevalence of llms in the scholarly literature. *ArXiv*, abs/2403.16887.

German Gritsai, Ildar Khabutdinov, and Andrey Grabovoy. 2024. Multi-head span-based detector for AI-generated fragments in scientific papers. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 220–225, Bangkok, Thailand. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics:*

EMNLP 2023, pages 7001–7025, Singapore. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucinations detections. *arXiv preprint*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Maksim Savkin, Anastasia Voznyuk, Fedor Ignatov, Anna Korzanova, Dmitry Karpov, Alexander Popov, and Vasily Konovalov. 2024. DeepPavlov 1.0: Your gateway to advanced NLP models backed by transformers and transfer learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 465–474, Miami, Florida, USA. Association for Computational Linguistics.

Jianbin Shen, Junyu Xuan, and Christy Liang. 2023. Mitigating intrinsic named entity-related hallucinations of abstractive text summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15807–15824, Singapore. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weihang Su, Yichen Tang, Qingyao Ai, Changyue Wang, Zhijing Wu, and Yiqun Liu. 2024. Mitigating entity-level hallucination in large language models.

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.

Amelie Wuehrl, Lara Grimminger, and Roman Klinger. 2023. An entity-based claim extraction pipeline for real-world biomedical fact-checking. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 29–37, Dubrovnik, Croatia. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Forty-first International Conference on Machine Learning*.

A Analysis of Provided Annotations

See Table 2 for examples of inconsistencies within annotations.

Id	Question	Answer
2	In which city was **David Sandberg** born?	David Sandburg was born in Stockholm, Sweden.
10	Is there a limit to the number of terms one can serve as the **Columbian president**?	Yes, there is a term limit for the Colombian president. The current president, Juan Manuel Santos, is serving his second and final term, which ends in 2020.
12	Are there any plants of the genus **Abrus** native to the Americas?	Yes, there are several species of Abrustum native in the American continent. Some of them include Abrutilis, Abrunosa, and Abrutus.
32	Explain in a few words the **Air France Flight 296Q crash.**	Horrific mid-air collision.
43	How many people visit the **Groninger Museum** yearly?	I'm sorry, I do not have that information. However, the museum's website states that it has over 100,000 visitors per year.
55	Who was the **Pasteur lunar crater** named after?	The Pasteur crater on the Moon is not named after Louis Pasteur, the famous French microbiologist. Instead, it is named after the French scientist and science administrator, Camille Alphonse Grenet-Bleuzenet de La Vernardiere Pasteur...
62	Who is the mayor of **Mercy-le-Haut**?	I cannot provide an answer to that question as Mercy-le-Haut is a fictional town in the Belgian comic book series "The Adventures of Tintin," created by Hergé...
131	Which Olympic sport did **Gergely Kulcsár** compete in?	Gerd Kulaşar competed in the men's hammer throw at the 1992 Summer Olympics.
135	What was the population of **Palm River-Clair Mel** in 2010?	In 2010, Palm river-clair mel had a population is 1,974.
146	How many presidential candidates were there in the **2013 Iranian presidential election**?	There were three candidates: Ebrahim Raisi, Hasan Rowhani, and Mohammad Bagher Ghalibaf.

Table 2: Some examples from the test set and comparison of spans from our annotation and ground truth annotation. Blue text corresponds to our spans, Red text correspond to annotators spans and Purple text to overlapping spans. Text in bold is the main entity of the question.

Id	Question	Answer
2	In which city was **David Sandberg** born?	David Sandburg was born in Stockholm, Sweden.
10	Is there a limit to the number of terms one can serve as the **Columbian president**?	Yes, there is a term limit for the Colombian president. The current president, Juan Manuel Santos, is serving his second and final term, which ends in 2020.
12	Are there any plants of the genus **Abrus** native to the Americas?	Yes, there are several species of Abrustum native in the American continent. Some of them include Abrutilis, Abrunosa, and Abrutus.
32	Explain in a few words the **Air France Flight 296Q crash.**	Horrific mid-air collision.
43	How many people visit the **Groninger Museum** yearly?	I'm sorry, I do not have that information. However, the museum's website states that it has over 100,000 visitors per year.
55	Who was the **Pasteur lunar crater** named after?	The Pasteur crater on the Moon is not named after Louis Pasteur, the famous French microbiologist. Instead, it is named after the French scientist and science administrator, Camille Alphonse Grenet-Bleuzenet de La Vernardiere Pasteur...
62	Who is the mayor of **Mercy-le-Haut**?	I cannot provide an answer to that question as Mercy-le-Haut is a fictional town in the Belgian comic book series "The Adventures of Tintin," created by Hergé...
131	Which Olympic sport did **Gergely Kulcsár** compete in?	Gerd Kulaşar competed in the men's hammer throw at the 1992 Summer Olympics.
135	What was the population of **Palm River-Clair Mel** in 2010?	In 2010, Palm river-clair mel had a population is 1,974.
146	How many presidential candidates were there in the **2013 Iranian presidential election**?	There were three candidates: Ebrahim Raisi, Hasan Rowhani, and Mohammad Bagher Ghalibaf.