FENJI at SemEval-2025 Task 3: Retrieval-Augmented Generation and Hallucination Span Detection

Flor Alberts, Ivo Bruinier, Nathalie de Palm, Justin Paetzelt, Erik Varecha
University of Groningen
[f.alberts.2, i.b.a.bruinier] @student.rug.nl
[n.h.m.de.palm, j.paetzelt, e.varecha] @student.rug.nl

Abstract

Large Language Models (LLMs) have significantly advanced Natural Language Processing, however, ensuring the factual reliability of these models remains a challenge, as they are prone to hallucination - generating text that appears coherent but contains innacurate or unsupported information. SemEval-2025 Mu-SHROOM focused on character-level hallucination detection in 14 languages. In this task, participants were required to pinpoint hallucinated spans in text generated by multiple instruction-tuned LLMs. Our team created a system that leveraged a Retrieval-Augmented Generation (RAG) approach and prompting a FLAN-T5 model to identify hallucination spans. Despite contradicting prior literature, our approach yielded disappointing results, underperforming all the "mark-all" baselines and failing to achieve competitive scores. Notably, removing RAG improved performance. The findings highlight that while RAG holds potential for hallucination detection, its effectiveness is heavily influenced by the retrieval component's context-awareness. Enhancing the RAG's ability to capture more comprehensive contextual information could improve performance across languages, making it a more reliable tool for identifying hallucination spans.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has significantly transformed Natural Language Processing (NLP), pushing breakthroughs in text generation, reasoning, and contextual understanding (Wang et al., 2024a). As these models continue to evolve, researchers have explored their potential across various domains, yet some challenges persist in ensuring the reliability and factual accuracy of their outputs (Ji et al., 2023).

A significant challenge in assessing LLM output is the phenomenon of hallucination, where models produce text that appears coherent but contains factually incorrect or unsupported information (Farquhar et al., 2024). This issue can stem from limitations in training data (McKenna et al., 2023), overgeneralization (Zhang et al., 2024), and the tendency of models to prioritize linguistic fluency over factual accuracy (Wang et al., 2024b). Existing evaluation metrics often focus on grammaticality and coherence, which is not able to properly account for, and penalize factual inconsistencies, making hallucinations more common (Honovich et al., 2022). Addressing this challenge is important for applications such as automated knowledge retrieval (Shi et al., 2025), decision support systems (Handler et al., 2024), and scientific content generation (Rossi et al., 2024), where misinformation can lead to potential consequences (Rawte et al., 2023; Asgari et al., 2024).

In a collaborative effort to develop the field of mitigating LLM hallucinations, the SemEval-2025 Mu-SHROOM shared task focuses on detecting hallucinated spans in text generated by instruction-tuned LLMs across multiple languages (Vázquez et al., 2025). Unlike its previous iteration, this task focuses on character-level hallucination detection in 14 different languages. Participants were given LLM-generated text, produced by multiple LLMs, and had to identify hallucinated characters while assigning confidence scores to their predictions. Evaluation was based on intersection-over-union (IoU) accuracy and the correlation between assigned probabilities and empirical annotations.

To approach this task, our team used a RAG approach for passage retrieval and the prompting of a FLAN-T5 model (Chung et al., 2022) as a method to detect spans of hallucinations. This method relied on using relevant and factually correct passages to be given to the T5 model, then leveraging its abilities to specifically identify what parts of a given piece of text could be a hallucination with a probability estimate. While our experiments showed middling results, it provides promis-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1143–1151

ing insight into using RAG as a tool for detecting hallucination spans. Our evaluations across 14 languages indicate that while the RAG component sometimes aids in pinpointing hallucinated spans, it often falls short. Our findings offer practical insights into further refining retrieval-augmented methods in hallucination detection.

2 Background

As per the definition provided by the Mu-SHROOM organizers, hallucinations are understood as content that contains or describes facts that are not supported by the provided reference (Vázquez et al., 2025). Broadly, hallucinations in LLMs can be classified into intrinsic and extrinsic hallucinations. Intrinsic hallucinations arise when some generated text is inconsistent with the input or reference material, introducing inaccuracies even when the model remains within its contextual boundaries. On the other hand, extrinsic hallucinations occur when a model produces information that extends beyond the provided context, fabricating unsupported claims (Ji et al., 2023; Wang et al., 2024c). In the context of the Mu-SHROOM task, detection of hallucination spans must be able to specifically identify intrinsic hallucinations, as extrinsic hallucinations do not fall under the definition of the task.

Although several methods have been explored for handling hallucinations in LLM output (Sanyal et al., 2024; Zhang et al., 2025), one notable method is RAG, which integrates external knowledge sources into LLM generation to improve factual consistency (Ayala and Bechard, 2024). This is typically implemented through a neural retriever, which retrieves relevant passages from a structured dataset (Lewis et al., 2020). Unlike traditional sparse retrieval methods like BM25, which rely on keyword matching, neural retrievers use dense embeddings to capture semantic relationships, which should improve retrieval accuracy (Lewis et al., 2021). By incorporating external knowledge retrieval, RAG has been shown to improve factual accuracy in NLP tasks as it reduces hallucinations by grounding responses in verifiable sources (Ayala and Bechard, 2024; Reichman and Heck, 2024; Karpukhin et al., 2020).

A key component that enhances RAG's retrieval process is Dense Passage Retrieval (DPR), which is a technique that uses dense vector representations to index and retrieve relevant passages for a given input. DPR uses a dual-encoder framework, where one encoder processes the input query while another encoder retrieves semantically similar documents. This allows DPR to efficiently retrieve top-k passages, which are then given to the RAG model for more context-aware generation (Karpukhin et al., 2020). Although more traditional methods could be effectively employed for RAG applications (Huly et al., 2024), by retrieving high-quality relevant passages, DPR has been shown to improve the factual reliability of RAG-based models (Lee and Kim, 2024).

The model focused on in this study, FLAN-T5, is a fine-tuned variant of the T5 model trained on diverse instruction-following tasks, and it is well-suited for applications that require contextual consistency and fact verification (Chung et al., 2022; Guan et al., 2024). Its ability to generalize across unseen tasks makes it particularly effective for detecting semantic inconsistencies in generated texts, which is a great benefit in hallucination detection. Since FLAN-T5 works in a text-to-text format, it could also be prompted to extract hallucinated spans directly, making it a promising tool for fine-grained hallucination detection. In last years SHROOM task, a study fine-tuned a FLAN-T5 for definition modeling, where it achieved an accuracy of 72.4% in detecting inconsistencies between input and generated definitons, demonstrating its potential for hallucination detection (Griogoriadou et al., 2024).

3 System Overview

The implementation of our system is publicly available on GitHub [1]. Figure 1 shows the pipeline of our system.

3.1 Data description

The data is provided in JSONL format, where each line corresponds to a single data entry structured in JSON. Each entry contains the prompt given to the language model and the generated output. Additionally, the model id is included for each entry. In the validation data, two additional types of annotations are included, which are soft and hard labels indicating hallucinations. The soft labels provide token spans (start and end indices) with an associated probability, which is calculated based on annotator agreement. Hard labels are a binary subset of these spans, derived by including only

[1] https://github.com/ivobruinier1/mu-SHROOM.git

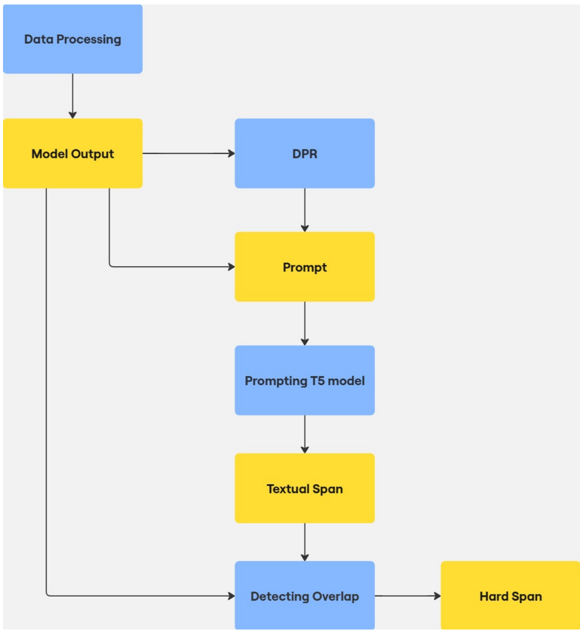

Figure 1: Pipeline for Extracting Hard Spans from Model Outputs

those soft labels with prob values above a threshold of 0.5. In both the training and test data, model output logits as well as model output tokens are provided.

For clarity, an example line from each subsection of the dataset is provided in Appendix A, where the structure and annotations can be examined in detail.

3.2 Dense Passage Retrieval

In our effort to optimize our prompt-based approach to detect hallucination spans, we leverage Dense Passage Retrieval (DPR) to provide context to the model. We aim to utilize this process in a manner that balances accuracy and performance to ensure usability in real world scenarios. To achieve this, we adopt a three-step approach for retrieving relevant passages.

After inspection of the training and validation data, we note that to answer most questions correctly, we would need to access domain-specific knowledge to some extent. For example, answering the question "*Do all arthropods have antennae?*" requires us to know the specific characteristics of arthropods. Based on this assumption, we implement Named Entity Recognition (NER) to extract named entities from each input query. For each of these entities, we search for the most likely Wikipedia pages using the Python Wikipedia module [2], which we then split into shorter passages. In our pipeline, we leverage multilingual transformer-based NER models to ensure optimal accuracy. Four distinct models are used, namely *roberta-ner-multilingual* [3] (Schelb et al., 2022), *robeczech-NER* [4] (trained using the *robeczech-base* model by Straka et al. (2021)), berteus-base-cased [5] (Agerri et al., 2020) and finbert-ner [6].

The second step in our pipeline involves the generation of more concise passages that can be used to provide context to the T5 model. After retrieving the most likely Wikipedia pages for each relevant entity, we split each page into sections of at most 5 sentences. Each section shares two sentences that overlap with the previous section in an attempt to retain context as much as possible.

As a final step in our DPR pipeline, we perform a semantic search where we compare each query in the test data to each passage relevant to the query. To achieve this, we implement a dual-encoder framework; we embed all passages for each query into a 384-dimensional dense vector space using *Sentence-BERT* [7] (Reimers and Gurevych, 2019). We then encode each input query using the same procedure. Finally, we retrieve the top-k=5 passages that are most relevant to the query to pass as context in the RAG prompt.

The language support for each individual NER model is shown in appendix C. As displayed here, none of these models offer support for Swedish and Farsi. As a workaround, we instead rely on Cohere Embed v3 [8] to perform a semantic search for both of these languages; however, due to computational cost and time constraints, we limit the number of included passages to the first 1,000,000 results.

3.3 T5 Span Detection

In our pipeline, the T5 model (google/flan-t5-base) is utilized to detect hallucination spans within the generated text. The process begins by reading data from JSONL files, which include model outputs and corresponding passages retrieved by a DPR system. The data is combined into pairs for further processing. Prompts are then generated using this

[2] https://pypi.org/project/wikipedia/
[3] https://huggingface.co/julian-schelb/roberta-ner-multilingual
[4] https://huggingface.co/popelucha/robeczech-NER
[5] https://huggingface.co/ixa-ehu/berteus-base-cased
[6] https://huggingface.co/Kansallisarkisto/finbert-ner
[7] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
[8] https://cohere.com/blog/introducing-embed-v3

data, which include context and hypotheses, and are formatted to query the T5 model. The model generates outputs based on these prompts, identifying potential hallucinations. To find the longest contiguous overlapping span between the T5 output and the text that could contain hallucinations, a sequence matching system is used. This involves preprocessing both texts by converting them to lowercase, removing punctuation, and normalizing whitespace to ensure consistent comparison. Python's SequenceMatcher (Python, 2025) is then applied to detect the longest common substring between the two inputs. The algorithm determines the start index and length of the best matching substring within the first text. If a valid overlap is found, the function returns the start and end indices of the match. If no overlap is detected, the function returns None. This method enables efficient detection of exact matches while ignoring variations in punctuation and capitalization, although it does not account for semantic similarity or minor textual differences, which could affect the precision of the span detection.

3.4 Evaluation

For evaluation the SemEval organizers released a scoring system [9] that could be implemented for reference and development of the system. The evaluation of intersection-over-union (IoU) of characters marked as hallucinations has been incorporated as way of providing feedback on how well the system scores. Our analysis does not include an evaluation of the correlation between the probability assigned by our system to a character being part of a hallucination and the empirical probabilities observed by the annotators. This decision was made due to limitations in the scope of the study. Future research may explore this aspect to better understand the alignment between automated predictions and human judgment.

$$\text{IoU} = \frac{\text{area of overlap}}{\text{area of union}} \tag{1}$$

4 Experiments & Results

4.1 Experimental setup

Experiments with the validation set were conducted using various prompting templates to evaluate their effectiveness. Multiple prompt variations were tested to determine which yielded the best performance. The most effective prompt template, as can be seen in the appendix 2, was then selected for the test set, where it was run both with and without DPR to assess the impact of retrieval augmentation on the results.

[9]https://github.com/Helsinki-NLP/shroom.git

4.2 Results and Discussion

Language	**IoU** FLAN-T5	**IoU** FLAN-T5 + DPR	**IoU** Baseline*
Arabic	0.00	0.05	0.36
Catalan	0.18	0.15	0.24
Czech	0.11	0.05	0.26
German	0.16	0.12	0.35
English	0.19	0.15	0.35
Spanish	0.13	0.13	0.19
Basque	0.13	0.13	0.37
Farsi	0.00	0.00	0.20
Finnish	0.09	0.07	0.49
French	0.08	0.08	0.45
Hindi	0.00	0.00	0.27
Italian	**0.23**	**0.28**	0.28
Swedish	0.12	0.09	0.54
Chinese	0.00	0.04	0.48

Table 1: IoU scores for all languages on the test data with the baseline (mark all)* scores for comparison

Previous studies (Ayala and Bechard, 2024; Reichman and Heck, 2024; Karpukhin et al., 2020) prove that RAG demonstrates potential in improving generative model performance. However, when the information retrieved by the DPR component is overly general or insufficiently relevant to the query, it can mislead the generative model, impairing its ability to accurately identify hallucinations. The study of Wu et al. (2022) highlights this by noting that passages often consist of multiple sentences, each potentially addressing different topics. Modeling such a passage as a single dense vector can be suboptimal. Error analysis of our results confirm the study of Wu et al. (2022), as DPR often retrieved information directly related to specific noun phrases but failed to capture information pertaining to the overall context of the entire sentence. This limitation has lead to incomplete or less relevant retrieval results which correlates to the lower IoU scores for most tested languages as can been seen in Table 1. Here, we observe that FLAN-T5 alone struggles across many languages, with scores of 0.00 or slightly higher for Arabic, Farsi, Hindi, and Chinese. Adding DPR seems to offers minor improvements in some cases, such as Arabic and Chinese, providing increases of 0.05 and 0.04, respectively. However, for several languages, like Czech and Catalan, combining FLAN-T5 with DPR leads to a decrease in IoU compared to using FLAN-T5 alone. For Farsi, the IoU remains the same at 0.00. For all languages, our FLAN-T5 setup fails to improve on the baseline scores. For the FLAN-T5 with DPR setup, Italian stands out as an exception, as it achieves an IoU identical to the baseline (0.28). This shows that only when testing the Italian dataset for hallucinations the DPR component was beneficial. Notably, scores for Italian are consistently high across all participating systems, indicating that the task may be inherently easier in Italian rather than reflecting an intrinsic advantage of this specific system for the language.

Additionally, the system's performance falls below the "mark all" baseline across all evaluated languages. Error analysis further supports the conclusion that a more generous span detection strategy could have led to improved results. However, when changing the prompt template in Figure 2 to be more generous, the system failed to achieve competing results.

4.3 Error Analysis

When analyzing the English textual output of the FLAN-T5 model, its performance varied. The model sometimes accurately detected hallucinations and maintained strong alignment with the content. However, it struggled with identifying hallucinations in long and complex outputs. FLAN-T5 was unable to produce multiple spans and often failed to label any hallucination at all. Additionally, information loss occurred during the conversion of FLAN-T5's output into hard labels, particularly due to the overlap detection segment of the system. Even when the model successfully identified hallucinations, some details were lost in the hard labeling process. As a result, the system's overall scores remained low. Notably, the model performed best when detecting hallucinations involving names of people or places. An example can be found in the appendix as Table 2.

4.4 Limitations

Languages that use non-alphabetical characters, such as Arabic, Farsi, and Chinese, do not perform well with this system. However, the FLAN-T5 + DPR system still attempted to detect some spans, suggesting that it is not entirely incapable of processing these languages, though its effectiveness is limited. The basis for this observation could be the model's tokenization and embedding process, which may not be well-suited for non-alphabetical scripts. Notably, overlap detection was minimal, indicating that the model struggled to correctly identify shared spans. Improving overlap detection could have led to better overall scores by enhancing the system's ability to capture relevant spans more accurately. For example, The overlap detection did not account for semantic similarity or minor textual differences which could have significantly affected the precision of the span detection. Furthermore, FLAN-T5 nor the overlap detection were able to capture multiple hallucination spans, outputting only a single span of hard labels for each detected hallucination. This limitation led to inaccurate detection, particularly when hallucinations were distributed across different parts of the text. Next to that, this study focused solely on the use of the FLAN-T5 model and did not explore other models that might have been more effective for hallucination span detection. Examining alternatives, such as GPT-style models or other instruction-tuned architectures, could have provided a more comprehensive evaluation of the system's approach.

5 Conclusion

This study explored hallucination span detection as part of the SemEval-2025 Mu-SHROOM task using RAG with the FLAN-T5 model. This approach integrated DPR with generative capabilities to identify hallucination spans. However, the system underperformed across all languages compared to the "mark-all" baselines. Notably, the removal of the RAG component led to improved performance, highlighting fundamental challenges with the retrieval mechanism's contextual relevance. The findings underscore the importance of robust retrieval mechanisms that can capture comprehensive contextual information. Future work could explore using different generative models, running detailed tests on the parts of the RAG system, and studying how language differences affect performance. Improving overlap detection could also help the system better identify hallucination spans. By working on these areas, RAG could become a more reliable method for detecting hallucinations.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, and Dominic Pimenta. 2024. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *medRxiv*.

Orlando Ayala and Patrice Bechard. 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, page 228–238. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Natalia Griogoriadou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. 2024. Ails-ntua at semeval-2024 task 6: Efficient model tuning for hallucination detection and analysis. *arXiv preprint arXiv:2404.01210*.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. *Preprint*, arXiv:2310.14564.

Abram Handler, Kai R. Larsen, and Richard Hackathorn. 2024. Large language models present new questions for decision support. *International Journal of Information Management*, 79:102811.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *Preprint*, arXiv:2204.04991.

Oz Huly, Idan Pogrebinsky, David Carmel, Oren Kurland, and Yoelle Maarek. 2024. Old ir methods meet rag. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2559–2563, New York, NY, USA. Association for Computing Machinery.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yanlin Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys (CSUR)*, 55(12):1–38.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Juhwan Lee and Jisu Kim. 2024. Control token with dense passage retrieval. *Preprint*, arXiv:2405.13008.

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.

Patrick 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. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *Preprint*, arXiv:2305.14552.

Foundation Python, Software. 2025. *difflib — Helpers for computing deltas*. Available at: `https://docs.python.org/3/library/difflib.html` [Accessed: 2025-02-12].

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S. M Towhidul Islam Tonmoy, Aman Chadha, Amit P. Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models – an extensive definition, quantification, and prescriptive remediations. *Preprint*, arXiv:2310.04988.

Benjamin Reichman and Larry Heck. 2024. Retrieval-augmented generation: Is dense passage retrieval retrieving? *arXiv preprint arXiv:2402.11035*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Luca Rossi, Katherine Harrison, and Irina Shklovski. 2024. The problems of llm-generated data in social science research. *Sociologica*, 18(2).

Soumya Sanyal, Tianyi Xiao, Jiacheng Liu, Wenya Wang, and Xiang Ren. 2024. Minds versus machines: Rethinking entailment verification with language models. *arXiv preprint arXiv:2402.03686*.

Julian Schelb, Maud Ehrmann, Matteo Romanello, and Andreas Spitz. 2022. ECCE: Entity-centric Corpus Exploration Using Contextual Implicit Networks. In *Companion Proceedings of the Web Conference 2022*.

Xiang Shi, Jiawei Liu, Yinpeng Liu, Qikai Cheng, and Wei Lu. 2025. Know where to go: Make llm a relevant, responsible, and trustworthy searchers. *Decision Support Systems*, 188:114354.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

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.

Chen Wang, Jin Zhao, and Jiaqi Gong. 2024a. A survey on large language models from concept to implementation. *arXiv preprint arXiv:2403.18969*.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024b. Factuality of large language models in the year 2024. *arXiv preprint arXiv:2402.02420*.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024c. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*.

Jian Wu, Junwei Xie, Bing Tang, Hongyan Yan, and Lijun Wang. 2022. Dense passage retrieval: A study of the representations of multiple sentences in a passage. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*.

Bohan Zhang, Xiaokang Zhang, Jing Zhang, Jifan Yu, Sijia Luo, and Jie Tang. 2025. Cot-based synthesizer: Enhancing llm performance through answer synthesis. *Preprint*, arXiv:2501.01668.

Yuji Zhang, Sha Li, Jiateng Liu, Pengfei Yu, Yi R. Fung, Jing Li, Manling Li, and Heng Ji. 2024. Knowledge overshadowing causes amalgamated hallucination in large language models. *Preprint*, arXiv:2407.08039.

A Appendix: data structure examples

A.1 Training data


```
{"lang":"EN","model_id":"
   togethercomputer\/Pythia-Chat-
   Base-7B","model_input":"Do all
    arthropods have antennae?","
   model_output_text":" Yes, all
   insects and arachnids (
   including spiders, scorpions,
   and ticks) have at least one
   antenna. ","
   model_output_logits
   ":[-2.57427001, ...
   ,14.8357839584],"
   model_output_tokens":["\
   u0120Yes", ... ,"\u010a"]}
```


A.2 Validation data


```
{"id":"val-en-1","lang":"EN","
   model_input":"What did Petra
   van Staveren win a gold medal
   for?","model_output_text":"
   Petra van Stoveren won a
   silver medal in the 2008
   Summer Olympics in Beijing,
   China.","model_id":"tiiuae\/
   falcon-7b-instruct","
   soft_labels":[{"start":10,"
   prob":0.2,"end":12}, ..., {"
   start":69,"prob":0.9,"end
   ":83}],"hard_labels":[[25,31],
    ... ,[69,83]]}
```


A.3 Test data (unlabeled)


```
{"id":"tst-en-1","lang":"EN","
   model_input":"Did Alberto
   Fouillioux ever play in a
   world cup championship?","
   model_output_text":" No,
   Albero Foulois was not in any
   of the FIFA World Cup finals.\
   n","model_id":"
   togethercomputer\/Pythia-Chat-
   Base-7B","model_output_tokens
   ":["\u0120No", ... ,"\u010a
   "],"model_output_logits
   ":[-2.2868447304, ...
   ,14.6126556396]}
```


A.4 Prompt Template

```
prompt = (
    f"Question = {question}.
       Hypothesis = {hypothesis}.
        "
    f"Using this context: {
       context}, identify and
       cite the EXACT part of the
        hypothesis that "
    f"contradicts the premise by
       giving the textual span,
       make sure to not add ANY
       other words."
)
```

Figure 2: Template used for prompting the FLAN-T5 model to identify hallucinations.

B Textual Error Analysis Example

tst-en-10	Textual Span
Model Input	"Is there a limit to the number of terms one can serve as the Columbian president?"
Model Output	"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."
Correct Spans	[current], [Juan Manuel Santos], [is serving], [second and final], [which ends in 2020]
FLAN-T5 Span	[Juan Manuel Santos is serving his second and final term, which ends in 2020.]
FENJI Span	[t, Juan Manuel Santos, is serving his second and final term, which ends i]

Table 2: Example of Textual Error Analysis for Datapoint tst-en-10: Demonstrating Information Loss and the Model's Inability to Detect Multiple Spans

C NER language support

Model	Languages
roberta-ner-multilingual	DE, EN, ES, ZH, CC, FR, AR, IT, HI
robeczech-NER	CS
berteus-base-cased	EU
finbert-ner	FI
Not Supported	FA, SV

Table 3: Language support for each NER model.