Hallucination Detectives at SemEval-2025 Task 3: Span-Level Hallucination Detection for LLM-Generated Answers

Passant Elchafei
Ulm University, Germany
passant.elchafei@uni-ulm.de

Mervat Abu-Elkheir
German University in Cairo, Egypt
mervat.abuelkheir@guc.edu.eg

Abstract

Detecting spans of hallucination in LLM-generated answers is crucial for improving factual consistency. This paper presents a span-level hallucination detection framework for the SemEval-2025 Shared Task, focusing on English and Arabic texts. Our approach integrates Semantic Role Labeling (SRL) to decompose the answer into atomic roles, which are then compared with a retrieved reference context obtained via question-based LLM prompting. Using a DeBERTa-based textual entailment model, we evaluate each role's semantic alignment with the retrieved context. The entailment scores are further refined through token-level confidence measures derived from output logits, and the combined scores are used to detect hallucinated spans. Experiments on the Mu-SHROOM dataset demonstrate competitive performance. Additionally, hallucinated spans have been verified through fact-checking by prompting GPT-4 and LLaMA. Our findings contribute to improving hallucination detection in LLM-generated responses.

1 Introduction

LLMs have demonstrated remarkable capabilities in generating human-like text, enabling a wide range of applications, including question-answering, summarization, and conversational agents. However, these models often produce hallucinations that appear plausible but are factually incorrect or unsupported by the input context (Quevedo et al., 2024).

Current research efforts focus on different approaches to detect and mitigate hallucinations. Some methods rely on sentence-level classification, where the entire response is labeled factual or not (Fadeeva et al., 2024). While these techniques provide a general assessment, they lack the granularity to pinpoint the specific hallucinated segments, leading to challenges in localized error correction (Liu et al., 2021). Other works explore token-level approaches, which utilize features such as minimum and average token probabilities to identify hallucinated content (Luo et al., 2024). This strategy shows promising results, but may struggle to effectively combine probabilistic and semantic information. These issues become more pronounced in morphologically rich and linguistically diverse languages such as Arabic, where complex syntactic rules and dialectal variations add layers of difficulty (Wang et al., 2023). Recent research highlights the need for improved hallucination detection techniques tailored to Arabic and other low-resource languages (Mubarak et al., 2024).

Our research introduces a span-level hallucination detection framework to address these challenges. Unlike previous sentence-level or token-level approaches, our framework identifies hallucinated spans by decomposing LLM-generated answers into semantic units using Semantic Role Labeling (SRL) and dependency parsing. These units are evaluated against context retrieved using GPT-4, then contradiction detection using pretrained textual entailment BERT model. Simultaneously, token-level confidence scores are computed to capture the model's certainty for each unit. The combined scores provide a refined measure of factuality which indicates hallucination parts. We evaluate our framework for both English and Arabic languages using the multilingual Mu-SHROOM dataset as part of the SemEval-2025 shared task. To further strengthen reliability, we incorporate an LLM-based verification step by using fact-checking technique using 2 different LLMs (GPT-4, and LLaMA).

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 introduces the dataset's structure. Section 4 presents our methodology. Section 5 outlines the experimental results. Finally, Section 6 concludes with key findings and future work.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 601–606

2 Related Work

Hallucination detection in LLMs has been studied, with research mainly focusing on sentence-level classification. These approaches determine whether an entire response is factual or hallucinated, but lack the granularity to identify specific hallucinated spans. Token-based methods leverage confidence measures to estimate factuality, but often fail to account for semantic inconsistencies within generated responses (Wang et al., 2023). Recent advances have emphasized the need for hallucination detection at the spectral level, allowing a finer-grained factual assessment (Ji et al., 2023).

Reference-based methods compare generated text against external sources such as Factcheck-Bench (Qiu et al., 2023), a fine-grained fact-checking benchmark, and HALoGen (Ravichander et al.), a large-scale hallucination evaluation suite that categorizes hallucination errors. Other techniques, such as InterrogateLLM (Varshney et al.), employ self-consistency verification, where LLMs are prompted multiple times to detect contradictions in their responses. Although these methods improve factual verification, they operate primarily at the response level rather than identifying hallucinated spans.

Textual entailment models have also been explored for hallucination detection, classifying responses into entailment, contradiction, or neutrality (Wadden et al., 2022). However, these approaches are typically sentence-level, limiting their effectiveness for pinpointing hallucinated spans(Chen et al., 2023). The detection of low resources languages hallucinations presents additional challenges, particularly in morphologically rich languages like Arabic, where syntactic complexity and dialectal variations complicate the factual verification (Senator et al., 2025). Datasets such as Halwasa (Mubarak et al., 2024) and ACQAD (Sidhoum et al., 2022) were developed to help in analyzing factual and linguistic inaccuracies. Research on hallucination detection in Arabic has largely focused on sentence-level classification, leaving a gap in span-level hallucination identification (Abdelazim et al., 2024). Some studies have introduced dependency parsing techniques and Semantic Role Labeling (SRL) to improve hallucination detection which highlight the role of syntactic decomposition in improving the evaluation of facts (Liu et al., 2023).

Advances in SRL for information extraction should also contributed to hallucination detection, particularly in low-resource language. Unlike traditional information extraction tasks, SRL-based techniques focus on verifying atomic claims within generated text, enhancing factual alignment with external sources. Although recent studies focus on hallucination detection, there remains a need for more linguistically adaptive approaches to improve the factual consistency in LLM-generated text.

3 Dataset Structure

The dataset used in this research is provided by the SemEval-2025 shared task, Mu-SHROOM (Multilingual Shared-task on Hallucinations and Related Observable Over generation Mistakes). This dataset contains examples in multiple languages, including English, Arabic, Spanish, and French. Each example in the dataset comprises a question-answer pair, the corresponding LLM-generated output, and annotations for hallucinated spans.

3.1 Dataset Schema

Each data entry in the dataset includes the following fields:

- ID: Unique identifier (e.g. "val-en-1").
- Language: Question and answer language ("EN" for English, “AR” for Arabic).
- Question: Question provided to the model (e.g. "What did Petra van Staveren win a gold medal for?", “كم مرحلة يتكون منها رجيم دوكان؟”).
- Answer: LLM-generated answer (e.g. "Petra van Stoveren won a silver medal in the 2008 Summer Olympics in Beijing, China.", "يعتمد المراحل في رجِيم دُوْكَان على النسخة المحددة من البرنامج ، لكن عادةً ما يتألف من خمسة مراحل.").
- Model Information: Model that generated the output (e.g., "tiiuae/falcon-7b-instruct", "openchat/openchat-3.5-0106-gemma").
- Soft Labels: Span-level annotations indicating parts of the text that may be hallucinated, with associated probabilities, e.g. "start": 10, "prob": 0.2, "end": 12.
- Hard Labels: Annotated spans confirmed as hallucinations, represented by fixed token positions (e.g. [25, 31], [45, 49]).
- Model Output Tokens: Tokens of the answer.
- Model Output Logits: List of logits (one per token), reflecting the confidence of the model for the prediction of each token. This is the field that is used to calculate the confidence score in our approach.

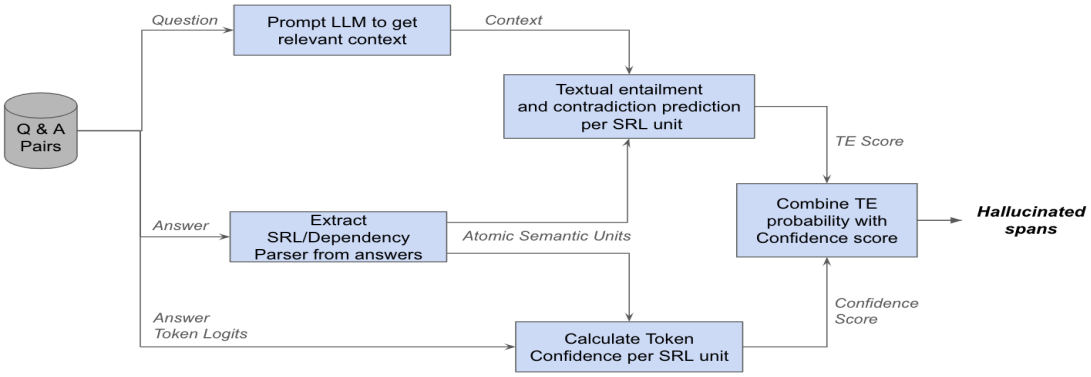

Figure 1: Span-Level Hallucination Detection Framework

4 Span-level Hallucination Detection Methodology

Our framework detects hallucinated spans in LLM-generated answers by combining semantic analysis with token-level probabilistic measures, as described in Figure 1. The system comprises several components: context retrieval for the question, answer decomposition, confidence scoring, textual entailment, score integration to get refined score which indicates the hallucination parts. The framework is designed to handle both English and Arabic, utilizing tools and models suitable for each language.

4.1 Context Retrieval

Given an input question, the relevant context is retrieved using the GPT-4 model with a well-written prompt. Separating retrieval from answer generation ensures an independent factual grounding mechanism. Moreover, retrieving relevant context is more reliable and less complex than generating a well-structured answer, as it relies on matching and ranking mechanisms, whereas answer generation requires reasoning, synthesis, and fluency while ensuring factual correctness. Based on this, the retrieved context serves as a factual reference for evaluating the generated answer and is expected to contain key facts necessary to comprehensively address the question. Additionally, considering the nature of the dataset, which consists of general knowledge question-answer pairs rather than specialized domain-specific queries, this approach is highly suitable. For example, for the question "What did Petra van Staveren win a gold medal for?" the retrieved context may include biographical details about her achievements in swimming, ensuring a factual basis for assessment.

4.2 Answer Decomposition

The generated answer is decomposed into atomic units using Semantic Role Labeling model (SRL). This step extracts structured components (e.g., predicates, arguments) that represent key facts. Each unit is later evaluated for factual alignment with the retrieved context. Example decomposition for "Petra van Staveren won a silver medal in the men's 10 km walk at the 2008 Summer Olympics":

- Verb: "won"
- ARG0: "Petra van Staveren"
- ARG1: "a silver medal"
- ARGM-TMP: "at the 2008 Summer Olympics"

The AllenNLP BERT-based SRL model (Shi and Lin, 2019) is used for English. For Arabic, we conduct two experiments: 1) HanLP's multilingual SRL extraction model, which supports Arabic language (He and Choi, 2021), with an example of an Arabic sentence shown in Figure 2. 2) CamelParser (Elshabrawy et al., 2023) for dependency parsing, followed by an SRL extraction algorithm.

CamelParser2.0 is an open-source Python-based Arabic dependency parser designed for the Columbia Arabic Treebank (CATiB) and Universal Dependencies (UD) formalisms. It processes raw text to perform tokenization, part-of-speech tagging, and morphological analysis, enhancing syntactic parsing, which is essential for accurate semantic role decomposition.

4.3 Confidence Scoring

Token-level confidence scores are computed using logits from the LLM output. These logits scores are given within the SemEval dataset for each token as described in the Dataset Structure section. A low score indicates reduced confidence, suggesting

potential hallucination. The logit score for each atomic unit is computed as described in Equation 1. Where n represents the total number of tokens per unit. $logit_i$ denotes the logit value of token i. The denominator $\sum_j e^{logit_j}$ represents the normalization factor across all tokens, ensuring that the computed probability is within the valid range [0,1]. The final logit score is the average softmax probability of tokens in the generated output.

$$logit\_score = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{logit_i}}{\sum_j e^{logit_j}} \quad (1)$$

This formulation effectively captures the confidence level, with lower scores indicating higher potential hallucination.

4.4 Textual Entailment

To evaluate factual alignment, each atomic unit is compared with the retrieved context, which is described in Section 4.1, using a natural language inference (NLI) model. We choose the DeBERTa (He et al., 2021) entailment model, which predicts whether the context entails, contradicts, or is neutral to the unit. This step generates a set of probabilities corresponding to the entailment, neutral, and contradiction labels. For instance, given the hypothesis "in the 2008 Summer Olympics in Beijing, China", the model output might be: Entailment: 1.1%, Neutral: 8.7%, Contradiction: 90.2%.

4.5 Score Integration

The entailment and confidence scores are combined to produce a refined score for each atomic unit. This score determines the likelihood that the unit is factual. A hyperparameter "α" controls the weight of the components of entailment and confidence as described in Equation 2.

$$refined\_score = \alpha \cdot entailment + (1-\alpha)\cdot confidence \quad (2)$$

5 Expriments and Results Analysis

Our experiments focus on two languages (English and Arabic) and assess performance using both intrinsic evaluation metrics and fact-checking verification with LLMs.

5.1 Evaluation Metrics

To measure the accuracy and effectiveness of hallucination detection, we evaluate our framework using Intersection over Union (IoU) and Correlation score (Cor) which are the used mertrics by the shared task (Vázquez et al., 2025). The IoU metric quantifies the overlap between predicted hallucinated spans and the ground truth annotations, where a higher value indicates improved span-level detection accuracy. The Correlation (Cor) metric evaluates the consistency between predicted hallucination probabilities and the ground truth confidence scores, providing insight into the model's reliability in detecting hallucinated spans.

5.2 Hallucination Identification

Units flagged with low refined scores are aggregated and reported as hallucinated spans, as shown in Figure 1. This enables fine-grained identification of factual inconsistencies within the answer text. Here is an example generated by our framework, showing that it is a "neutral" entailment but because of low logit score based on the output token logit given by the generated answers dataset. Refined Score threshold is set at 0.5 to identify hallucinated spans, serving as a balanced decision point. Units with a refined score below this value are classified as hallucinations.


```
"ARG1": {
    "hypothesis": "a silver medal",
    "predicted_label": "neutral",
    "entailment_probabilities": {
        "entailment": 0.7,
        "neutral": 75.3,
        "contradiction": 23.9
    },
    "logit_score": 0.3333333333333333,
    "refined_score": 0.1375,
    "hallucinated": true
}
```

```
"ARGM-LOC": {
    "hypothesis": "in the 2008 Summer
        Olympics in Beijing , China",
    "predicted_label": "contradiction",
    "entailment_probabilities": {
        "entailment": 1.1,
        "neutral": 8.7,
        "contradiction": 90.2
    },
    "logit_score": 0.1111111111111111,
    "refined_score": 0.051,
    "hallucinated": true
}
```


5.3 Experimental Results

English Language Results: Our model achieves an IoU of 0.358 and a Cor of 0.322. To further validate hallucinated spans, we use GPT-4 and LLaMA to verify detected hallucinations. GPT-4 matched 83% of the hallucinated spans, while LLaMA

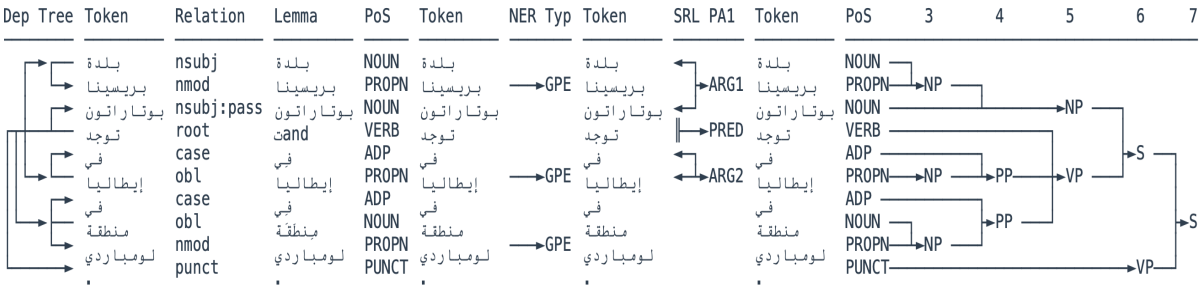

Figure 2: Example of Arabic sentence SRL extraction

identified 72%, demonstrating the effectiveness of LLM-based verification. GPT-4 achieved higher verification accuracy, successfully confirming or refuting hallucinated spans more consistently.

Arabic Language Results: The performance of our hallucination detection framework in Arabic was evaluated using two different models: 1) HanLP Multilingual model for SRL extraction and 2) CamelParser2.0 dependency parser model followed by SRL extraction algorithm. The HanLP model achieves an IoU of 0.205 and a Cor of 0.159, demonstrating lower performance compared to English. This discrepancy is attributed to the morphological complexity and syntactic variation in Arabic, which make span-level hallucination detection more challenging. An improvement is observed when employing CamelParser followed by SRL extraction algirthm, yielding an IoU of 0.28 and a Cor of 0.21. The increased accuracy suggests that syntactic parsing before semantic role decomposition provides better structured representations for hallucination detection. Additionally, the results indicate that this approach is more robust to dialectal variations, making it better suited for handling complex Arabic linguistic structures. Overall, the findings confirm that integrating dependency parsing improves hallucination detection in morphologically rich languages like Arabic. For the Arabic Fact-Checking Verification step, GPT-4 identified 58% of hallucinated spans, showing comparatively lower performance than in English. This is partly due to the GPT-4 model itself being less accurate with Arabic, particularly when handling ambiguous factual claims. However, we utilized GPT-4 to ensure a consistent evaluation approach between Arabic and English cases.

6 Conclusion

This paper presents a span-level hallucination detection framework that integrates Semantic Role Labeling (SRL), textual entailment, and token-level confidence scoring to identify hallucinations in LLM-generated answers. Evaluated on the MuSHROOM dataset, our approach achieves an IoU of 0.358 and a Cor of 0.322 in English, and an IoU of 0.28 and a Cor of 2.1 in Arabic when using CamelParser combined with the SRL extraction algorithm. Our results emphasize the effectiveness of dependency parsing before SRL extraction, particularly in Arabic, where linguistic complexity poses additional challenges. Despite these improvements, challenges remain for morphologically rich languages. Future work will focus on enhancing entailment models and addressing additional syntactic structures, such as nominal sentences in Arabic and long complex sentences.

References

Hazem Abdelazim, Tony Begemy, Ahmed Galal, Hala Sedki, and Ali Mohamed. 2024. Multi-hop arabic llm reasoning in complex qa. *Procedia Computer Science*, 244:66–75.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, et al. 2023. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*.

Ahmed Elshabrawy, Muhammed AbuOdeh, Go Inoue, and Nizar Habash. 2023. Camelparser2. 0: A state-of-the-art dependency parser for arabic. In *Proceedings of ArabicNLP 2023*, pages 170–180.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and

Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

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 Computing Surveys*, 55(12):1–38.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*.

Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. 2024. Hallucination detection and hallucination mitigation: An investigation. *arXiv preprint arXiv:2401.08358*.

Hamdy Mubarak, Hend Al-Khalifa, and Khaloud Suliman Alkhalefah. 2024. Halwasa: Quantify and analyze hallucinations in large language models: Arabic as a case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8008–8015.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. *arXiv preprint arXiv:2305.13632*.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. 2024. Detecting hallucinations in large language model generation: A token probability approach. *arXiv preprint arXiv:2405.19648*.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. The halogen benchmark: Fantastic llm hallucinations and where to find them.

Ferial Senator, Abdelaziz Lakhfif, Imene Zenbout, Hanane Boutouta, and Chahrazed Mediani. 2025. Leveraging chatgpt for enhancing arabic nlp: Application for semantic role labeling and cross-lingual annotation projection. *IEEE Access*, 13:3707–3725.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Abdellah Hamouda Sidhoum, M'hamed Mataoui, Faouzi Sebbak, and Kamel Smaïli. 2022. Acqad: a dataset for arabic complex question answering. In *International conference on cyber security, artificial inteligence and theoretical computer science*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by actively validating low-confidence generation.

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.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. Scifact-open: Towards open-domain scientific claim verification. *arXiv preprint arXiv:2210.13777*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.