HausaNLP at SemEval-2025 Task 3: Towards a Fine-Grained Model-Aware Hallucination Detection

Maryam Bala[1], **Amina Imam Abubakar**[1,2], **Abdulhamid Abubakar**[1], **Abdulkadir Shehu Bichi**[1], **Hafsa Kabir Ahmad**[1,3], **Sani Abdullahi Sani**[1], **Idris Abdulmumin**[1,4], **Shamsuddeen Hassan Muhamad**[1,3,5], **Ibrahim Said Ahmad**[1,3,6]

[1]HausaNLP, [2]University of Abuja, [3]Bayero University Kano, [4]Data Science for Social Impact, University of Pretoria, [5]Imperial College London, [6]Northeastern University

correspondence: maryam.bala@outlook.com, i.ahmad@northeastern.edu

Abstract

This paper presents our findings of the Multilingual Shared Task on Hallucinations and Related Observable Overgeneration Mistakes, MU-SHROOM, which focuses on identifying hallucinations and related overgeneration errors in large language models (LLMs). The shared task involves detecting specific text spans that constitute hallucinations in the outputs generated by LLMs in 14 languages. To address this task, we aim to provide a nuanced, model-aware understanding of hallucination occurrences in English. We used natural language inference and fine-tuned a ModernBERT model using a synthetic dataset of 400 samples, achieving an Intersection over Union (IoU) score of 0.032 and a correlation score of 0.422. These results indicate a moderately positive correlation between the model's confidence scores and the actual presence of hallucinations. The IoU score indicates that our model has a relatively low overlap between the predicted hallucination span and the truth annotation. The performance is unsurprising, given the intricate nature of hallucination detection. Hallucinations often manifest subtly, relying on context, making pinpointing their exact boundaries formidable.

1 Introduction

Despite the advancements in Natural Language Processing (NLP) and the development of Natural Language Generation (NLG) models, their limitations and potential risks have gained increased attention. A significant issue is that NLG models often produce unfaithful text relative to the source input, a phenomenon known as "hallucination" (Koehn and Knowles, 2017a; Rohrbach et al., 2018; Maynez et al., 2020a). Hallucinations in NLG models often result in outputs that, while fluent, lack accuracy. This issue arises because existing evaluation metrics prioritize fluency over correctness, ultimately diminishing system performance and failing to meet user expectations in practical applications (Mickus et al., 2024). For example, Dopierre et al. (2021) illustrate this phenomenon by attempting to paraphrase the statement "I am not sure where my phone is," which leads to the hallucinated output: "How can I find the location of any Android mobile."

Hallucination in NLG is concerning due to its impact on performance and safety in applications like medicine, where hallucinatory summaries or machine-translated instructions can pose risks to patient diagnosis (Ji et al., 2023). Similarly, hallucinations can lead to privacy violations by generating sensitive information not present in the source input (Carlini et al., 2021).

To address this challenge, *"SemEval-2025 Task 3 – Mu-SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes"* (Vázquez et al., 2025) was introduced. The shared task involves detecting specific text spans that constitute hallucinations in outputs generated by LLMs. Participants compute the probability of each character being marked as a hallucination using the LLM output consisting of a character string, tokens, and logits, thus enabling a fine-grained hallucination detection.

2 Background

Numerous efforts are provided to address hallucination across various NLG tasks. Analyzing hallucinatory content and its relationships in different tasks could enhance our understanding and unify efforts across NLG fields (Ji et al., 2023). While most existing studies focus on specific tasks like abstractive summarization (Huang et al., 2021; Maynez et al., 2020b) and machine translation (Lee et al., 2018), the study of Ji et al. (2023) offer a comprehensive analysis on the phenomenon of hallucination in abstractive summarization, dialogue generation, generative question answering, data- to-text generation, and machine translation.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1737–1741

(Parikh et al., 2020) argue that hallucination problem occurs when there is very little divergence in dataset and encoder with a defective comprehension ability could influence the degree of hallucination. Similarly, (Koehn and Knowles, 2017b) show that training and modeling choices of neural models have influence for hallucination. While large pre-trained models used for downstream NLG tasks are powerful in providing generalizability and coverage, they however prioritize parametric knowledge over the provided input and can result in hallucination of excess information in the output (Longpre et al., 2021).

Recent efforts to address hallucination include the Shared Task on Hallucinations and Related Observable Overgeneration Mistakes (SHROOM), which focused on binary classification for task-specific English language models (Mickus et al., 2024). Maksimov et al. (2024); Arzt et al. (2024); Rahimi et al. (2024) identify cases of fluent overgeneration hallucinations in model-aware and model-agnostic settings. They detect grammatically sound outputs which contain incorrect or unsupported semantic information. Building on this task instead of focusing on model-agnostic and model-aware tracks, this year's task focuses on the multilingual aspect. Therefore, all data-points in this year's task are model-aware.

We present our submission for the task *Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes (MU-SHROOM)* which aims to detect hallucinations in a multilingual context, providing a more nuanced understanding of their occurrence. MU-SHROOM is a SemEval-2025 shared task that focuses on detecting hallucinations and overgeneration errors in AI-generated text, a crucial challenge in improving the reliability of LLMs (Vázquez et al., 2025). Hallucinations occur when LLMs produces fluent but false or unsupported content, while overgeneration mistakes involve excessive, often misleading text (Ji et al., 2023). This task is important as language models become increasingly prevalent in various applications, where factual accuracy and reliability are essential for maintaining user trust and system integrity. Therefore, tackling these issues is essential for ensuring AI-generated text remains trustworthy and useful across various applications (Maksimov et al., 2024).

The task consists of participants detecting spans of text corresponding to hallucinations and determine which parts of the given text produced by LLMs constitute hallucinations. Annotated dataset are provided, allowing researchers to develop and benchmark models for identifying these issues in different linguistic contexts. Given the LLM output as a string of characters, a list of tokens, and a list of logits, participants calculate the probability that each character is marked as a hallucination and thus provide a fine- grained hallucination detection. Similarly, the task is held in multilingual and multi-model context as data are produced by a variety of public-weights LLM in multiple languages which includes: Arabic (Modern standard), Chinese (Mandarin), English, Finnish, French, German, Hindi, Italian, Spanish, and Swedish, along with three surprise test languages.

3 System Overview

One of the primary challenges in this task was the lack of labeled training data. To address this, we created a synthetic training dataset using ChatGPT, adhering to official annotation guidelines.

For the language selection, we used English language only and this is due to time constraint. Secondly, English language offers an extensive resource for NLP model development and evaluation. Similarly, focusing on a single language allowed us to develop deeper linguistic pattern recognition for hallucination detection.

For model selection, we fine-tuned ModernBERT (Warner et al., 2024), an advanced variant of BERT (Devlin et al., 2019) that offers significant improvements in handling long-context inputs, computational efficiency, and robustness in token classification tasks. This choice aligns with recent studies highlighting the benefits of long-context language models in detecting nuanced text errors (Sahitaj et al., 2025).

For a precise understanding of hallucination detection, we used the model-aware method. This method is based on analysing internal data of LLM during inference. One of the possible approaches is the analysis of the outputs of the hidden layers of the transformer. Using vector values of hidden layers for hallucination detection was proposed in a method called Statement Accuracy Prediction, based on Language Model Activations (SAPLMA) (Azaria and Mitchell, 2023). SAPLMA is a probing technique that utilises a feedforward neural network trained on activation values of the hidden

layers of LLM.

4 Experimental Setup

4.1 Dataset Generation

For this task, we initiated the development of a robust model for detecting hallucinations in text by generating synthetic data specifically for training purposes. Borra et al. (2024) have shown the success of using synthetic data in finetuning models for hallucination detection in LLMs. This synthetic dataset was designed to simulate a wide range of scenarios, enhancing the model's ability to generalize across various contexts. With this approach, We were able to generate 400 diverse labeled data points.

Similarly, we utilized the task-provided validation dataset to assess the model's performance during the training phase, ensuring that it effectively learned from the training data while maintaining its ability to generalize. Finally, we reserved the provided unlabeled test dataset for final evaluation, allowing us to measure the model's performance on completely unseen data.

4.2 Data Preprocessing

The text data was tokenized using the Hugging Face AutoTokenizer from the Transformers library (Wolf et al., 2020), which efficiently segmented the text into manageable units. Similarly, token labels were assigned based on entity spans identified in the dataset, ensuring that the model learned to recognize specific entities and their contexts.

4.3 Model Training

The core of our approach involved fine-tuning the ModernBert model on the synthetic dataset. During this training process, we implemented several advanced techniques to optimize performance:

- Cosine Learning Rate Scheduler: This scheduler dynamically adjust the learning rate throughout the training, promoting a smoother convergence and help avoid local minima.
- Mixed Precision Training: By utilizing both float16 and float32 data types, we enhanced computational efficiency while preserving model accuracy. This approach significantly reduced memory usage and improved training speed.
- Gradient Clipping: To maintain stability during training, we employed gradient clipping techniques that prevented gradients from exceeding a specified threshold. This safeguard helped mitigate issues related to exploding gradients, which can destabilize the training process.

4.4 Evaluation Metrics

To evaluate the model's effectiveness post-training, we utilized several key metrics: precision, recall, and F1 score. The selected evaluation metrics were implemented to facilitate a comprehensive assessment of the model's efficacy, encompassing both predictive precision and capacity to address class distribution asymmetries. The above metrics are used in addition to Intersection over Union (IoU) and Correlation Score that were used by the task organizers to assess the performance of our model.

4.5 Classification and Prediction

Once trained, the model was capable of processing unseen text to detect hallucinations effectively. The outputs of the model were converted into two formats: hard labels (binary classification) indicating the presence or absence of hallucinations, and soft labels representing confidence scores that quantify the model's certainty regarding its predictions. This dual-output approach not only enhances interpretability but also allows for flexible integration into downstream applications where varying levels of confidence may be required.

5 Results

At the end of training our model, the evaluation result showed that the model had a precision and recall score of 0.49 and 0.54 respectively. The F1 score of the model was at 0.43. The result indicates that the model correctly identifies hallucinations about half the time. The model is also able to detect slightly more than half of all hallucinations present in the text.

On the task-based evaluation of our model, our system achieved an Intersection over Union (IoU) score of 0.032 and a correlation score of 0.422. The IoU score indicates that our model has a relatively low overlap between the predicted hallucination span and the truth annotation. With the relatively low score, the model is struggling to identify the exact boundaries of the hallucinated content. Going by this result also, the model may be prone to false positives and/or false negatives.

The models correlation score of 0.422 shows a moderate positive correlation between the con-

fidence scores of the model and the actual presence of hallucinations. This result can translate to the model's ability to differentiate between hallucinated and non-hallucinated content. While this may not be the best performance, the results show that there is room for further improvement.

Our model results ranked 42nd and 24th on the IoU and correlation score indices respectively. The scores are not favourable and may not be entirely unexpected given the inherent complexity of hallucination detection,. Hallucinations can be subtle and context-dependent, making exact boundary detection particularly challenging. The model result is a promising starting point for further improvement.

6 Conclusion

This paper presents our approach to the SemEval shared task on LLM hallucination detection, focusing on a fine-grained, model-aware analysis of hallucination occurrences in the English language. We leveraged natural language inference and fine-tuned a ModernBERT model using a synthetic dataset of 400 samples. Our model achieved rankings of 42nd and 24th on the Intersection over Union (IoU) and correlation score indices, respectively. These results, while modest, underscore the inherent complexity of hallucination detection and highlights the need for continued refinement and innovation in this area. Our findings serve as a promising foundation for future improvements, emphasizing the importance of model-aware strategies in enhancing the reliability of LLM outputs.

References

Varvara Arzt, Mohammad Mahdi Azarbeik, Ilya Lasy, Tilman Kerl, and Gábor Recski. 2024. Tu wien at semeval-2024 task 6: Unifying model-agnostic and model-aware techniques for hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1183–1196.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Federico Borra, Claudio Savelli, Giacomo Rosso, Alkis Koudounas, and Flavio Giobergia. 2024. MALTO at SemEval-2024 task 6: Leveraging synthetic data for LLM hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1678–1684, Mexico City, Mexico. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. Protaugment: Unsupervised diverse short-texts paraphrasing for intent detection meta-learning. *arXiv preprint arXiv:2105.12995*.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

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.

Philipp Koehn and Rebecca Knowles. 2017a. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Philipp Koehn and Rebecca Knowles. 2017b. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020a. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020b. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Timothee Mickus, Elaine Zosa, Raúl Vázquez, 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.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Zahra Rahimi, Hamidreza Amirzadeh, Alireza Sohrabi, Zeinab Taghavi, and Hossein Sameti. 2024. Hallusafe at semeval-2024 task 6: An nli-based approach to make llms safer by better detecting hallucinations and overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 139–147.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Premtim Sahitaj, Iffat Maab, Junichi Yamagishi, Jawan Kolanowski, Sebastian Möller, and Vera Schmitt. 2025. Towards automated fact-checking of real-world claims: Exploring task formulation and assessment with llms.

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.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.