UMUTeam at SemEval-2025 Task 3: Detecting Hallucinations in Multilingual Texts Using Encoder-only Models Guided by Large Language Models

Ronghao Pan, Tomás Bernal-Beltrán, José Antonio García-Díaz, Rafael Valencia-García
Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{ronghao.pan, tomas.bernalb, joseantonio.garcia8, valencia}@um.es

Abstract

Large Language Models like GPT-4, LLaMa, Mistral, and Gemma have revolutionized Natural Language Processing, advancing language comprehension, generation, and reasoning. However, they also present challenges, particularly the tendency to hallucinate—that is, to produce false or fabricated information. This paper presents our participation in Task 3 MuSHROOM of SemEval 2025, which focuses on detecting hallucinations in multilingual contexts. Specifically, the task requires identifying text segments generated by LLMs that correspond to hallucinations and calculating the hallucination probability for each character in the text. To address this challenge, we adopted a token classification approach using the pre-trained XLM-RoBERTa-large model, fine-tuned on the provided training set. Additionally, we integrated context from Llama-3.1-70B to enhance hallucination detection by leveraging its broader and more up-to-date knowledge base. Our approach combines the multilingual capability of XLM-RoBERTa with the contextual understanding of Llama-3.1-70B, producing a detailed hallucination probability for each character in the text. The results demonstrate that our approach consistently outperforms baseline methods across multiple languages, particularly in detecting token-level hallucinations.

1 Introduction

Large Language Models (LLMs) such as GPT-4, LLaMa, Mistral, Gemma, and others, have marked a significant paradigm change in Natural Language Processing (NLP), achieving significant advances in language compression, generation, and reasoning (Brown et al., 2020). These models have overcome many limitations of previous models, enabling more robust and sophisticated applications in areas such as machine translation, virtual assistance, visual content generation, sentiment analysis, etc (Das et al., 2025).

One of the key factors of their success has been the massive scaling of parameters and training data, which has allowed them to capture linguistic nuances and broad domain knowledge. In addition, the development of more efficient and optimized architectures has improved their reasoning capabilities, allowing them to perform complex logic and analysis tasks.

Another revolutionary aspect has been the ability of LLMs to perform few-shot learning and even zero-shot learning, adapting to new tasks with few or no specific examples, expanding their versatility and applicability in dynamic environments. This has opened up new possibilities in areas such as customer service (Xiaoliang et al., 2024), personalized education (Wen et al., 2024), and AI-powered scientific research.

However, along with these remarkable advances, concerns have arisen about the tendency of LLMs to hallucinate, i.e., to produce false or fictitious information with high confidence and apparent consistency. These hallucinations can range from factually incorrect data to fabricated quotes or specific details that do not correspond to reality, posing significant risks in critical applications such as medicine, legal advice, or news broadcasting (Huang et al., 2025).

The root of the problem lies in the way LLMs generate text: based on statistical patterns learned from large amounts of data. While this allows them to produce fluid, contextually relevant responses, it also means that they have no real understanding of the world and no internal fact-checking (Guerreiro et al., 2023). As a result, they may combine information plausibly but incorrectly, leading to hallucinations. This complicates the practical implementation of LLMs, especially in real-world information retrieval systems that have become integrated into our daily lives, such as chatbots (Aljamaan et al., 2024), search engines (Shi et al., 2025), or content moderation (Pan et al., 2025, 2024; García-Díaz

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 750–756

et al., 2023).

It is important to note that hallucinations in conventional natural language generation (NLG) tasks have been extensively studied, and are defined as generated content that is illogical or not faithful to the provided source content. Moreover, in recent years, several tasks have been developed to detect hallucinations in LLMs, such as Shroom 2024 (Mickus et al., 2024), etc.

For this reason, Task 3 Mu-SHROOM of SemEval 2025 (Vázquez et al., 2025) was also released with the aim of detecting the parts of texts that correspond to hallucinations. The main goal of this task is therefore for participants to determine which parts of a text produced by LLM represent hallucinations. The task is organized in several languages and is performed in a multilingual context with different LLM. For each text, the probability that it is a hallucination must be generated for each character in the text.

To solve this task, we have adopted a token classification approach based on a pre-trained language model, specifically XLM-RoBERTa-large (Conneau et al., 2019), which is fine-tuned with the training set. However, in this case, for hallucination detection, the context, or response generated by an LLM, such as Llama-3.1-70B (Dubey et al., 2024), is added, as this model has broader knowledge. This improves hallucination detection by providing additional information that helps the model to distinguish between correct and hallucinatory content. Furthermore, the combination of XLM-RoBERTa-large with Llama-3.1-70B exploits the strengths of both models: the multilingual capability and robustness of XLM-RoBERTa, together with the up-to-date and contextual knowledge of Llama-3.1-70B. The fitted model generates the hallucination probability for each character in the text, providing a detailed and accurate analysis. This approach proves effective in a multilingual context and with different LLMs, optimizing hallucination detection in different scenarios.

2 Background

LLMs are large-scale language models designed to understand and process human language by learning contextual patterns and relationships in large volumes of textual data. These models have a large number of parameters and use advanced pre-training techniques, such as Masked Language Modeling (MLM) and autoregressive prediction, allowing them to accurately model probabilities and contextualized semantics of text (Huang et al., 2025).

In recent years, several high-profile LLMs have been developed, including OpenAI's GPT-4, Meta AI's Llama, Google's Gemma, Mixtral and Mistral. These models have demonstrated their versatility in a wide range of applications, from search engines and customer support to code generation, education, healthcare and financial analysis.

Despite their remarkable advances and versatility in multiple domains, LLMs present several limitations that affect their reliability and applicability in real environments, such as the generation of hallucinations, lack of fact and reasoning verification, context sensitivity, as well as biases and limitations in complex reasoning and specialized tasks. Therefore, there is a need for new approaches and more robust strategies to solve these problems and, thus, optimize resources in the future development of LLMs (Huang et al., 2025).

Thus, different techniques and approaches have emerged for detecting hallucinations in LLMs, such as fact-checking, which verifies the accuracy of generated content by cross-referencing it with reliable external sources (Min et al., 2023). Additionally, frameworks that utilize evidence gathering and internal verification tools leverage the knowledge stored within the LLMs themselves, using techniques like Chain-of-Thought (CoT) (Dhuliawala et al., 2024). However, these methods are limited, as LLMs are not always reliable data sources.

Several approaches attempt to detect hallucinations without relying on sources external to the LLMs themselves. These include analyzing internal signals such as token-level confidence scores or entropy measures (Varshney et al., 2023; Luo et al., 2024), as well as evaluating model behavior through consistency checks, for example by comparing multiple generations or using multi-agent discussion frameworks (Agrawal et al., 2024).

In our approach, we use a multilingual token classification model based on XLM-RoBERTa to identify hallucinated segments in text generated by LLMs. Instead of relying on external knowledge bases, we incorporate a reference response generated by Llama-3.1-70B as additional context. This setup allows the model to compare the original output with a high-quality alternative and detect inconsistencies at the token level. Our method uses internal information from the LLMs' outputs and comparison signals between model generations to

improve hallucination detection.

Specifically, XLM-RoBERTa is used as a token classification model to identify hallucinations generated by LLMs, incorporating the response generated by Llama-3.1-70B as context into the token classification input to improve detection accuracy. This integration allows the XLM-RoBERTa model to benefit from the enhanced generation and understanding capabilities of Llama-3.1-70B, increasing its effectiveness in identifying inconsistencies and errors in the output generated by other LLMs.

3 System overview

Figure 1 shows the system architecture, where the XLM-RoBERTa-Large model is used as the basis for performing fine-tuning in a token classification task. In this approach, each token in the text is classified with a value of 1 or 0, indicating whether that token corresponds to a hallucination (1) or not (0).

XLM-RoBERTa-Large is a multilingual version of RoBERTa, trained on 2.5 TB of CommonCrawl filtered data spanning 100 languages. This model was pre-trained using the Masked Language Modeling (MLM) objective, in which 15% of the words in the input are randomly masked, and the model must predict the masked words using the surrounding context.

Unlike traditional recurrent neural network models (RNNs), which process words sequentially, or autoregressive models such as GPT, which mask future tokens, XLM-RoBERTa leverages a bidirectional representation of the sentence. This allows it to learn an internal representation of 100 languages, which is essential for multilingual tasks.

To provide the model with a broader context or a pre-answer to the question, an LLM such as Llama-3.1-70B has been used. Since this model has been trained with a larger amount of data, it offers superior performance compared to the LLMs used to generate the dataset answers. To incorporate the response or context generated by Llama-3.1-70B, a `text_pair` configuration has been employed, in which the text generated by the model (`model_output_text`) and the context or response of the larger model (teacher model) are entered together.

This configuration allows XLM-RoBERTa-Large to consider not only the text itself, but also its context, which can help the model better distinguish between factual and hallucinated content.

During tokenization, labels are assigned to the tokens using the offset mappings generated by the tokenizer, which indicate the start and end positions of each token in the original text. Those tokens whose ranges match the positions specified in the hard labels, which mark the hallucinations in the text, are labeled as 1. All other tokens receive a label of 0.

This granular approach at the token level enables accurate and detailed detection of hallucinations, helping to identify exactly where in the text the hallucination occurs. In addition, by using XLM-RoBERTa-Large with text pairs (`text_pair`), a more robust and effective contextual analysis is achieved, maximizing model performance in a multilingual environment.

We used the HuggingFace transformers library with the XLM Roberta large model and its associated tokenizer. Each input sample was encoded using the `text_pair` format, where the first segment corresponds to the model-generated output (`model_output_text`) and the second segment corresponds to the context provided by Llama-3.1-70B model.

A token classification head with a single linear layer and softmax activation was added on top of the encoder. Although the output logits consist of three values per token (due to padding and special tokens), only two labels are used during training: 1 for hallucinated tokens and 0 for the rest. Tokens not aligned with any character (e.g., special tokens) were assigned a label of -100 to be ignored by the loss function.

The model was trained using the cross-entropy loss function, which is standard for token classification tasks. Offset mappings were used to align character-level annotations with token-level labels.

4 Experimental setup

For the experiment, the labeled training set provided by the organizers has been used with 50 examples are kept for each language, except for SV which has 49 examples.

In Figure 2, the length distribution of the input, the response generated by the LLM and the response generated by the LLM teacher (Llama-3.1-70B) is shown. It can be seen that the maximum length of the response generated by the LLM teacher is around 100 tokens, while the length of the responses generated by the LLM can reach up to about 500 tokens. This difference could be

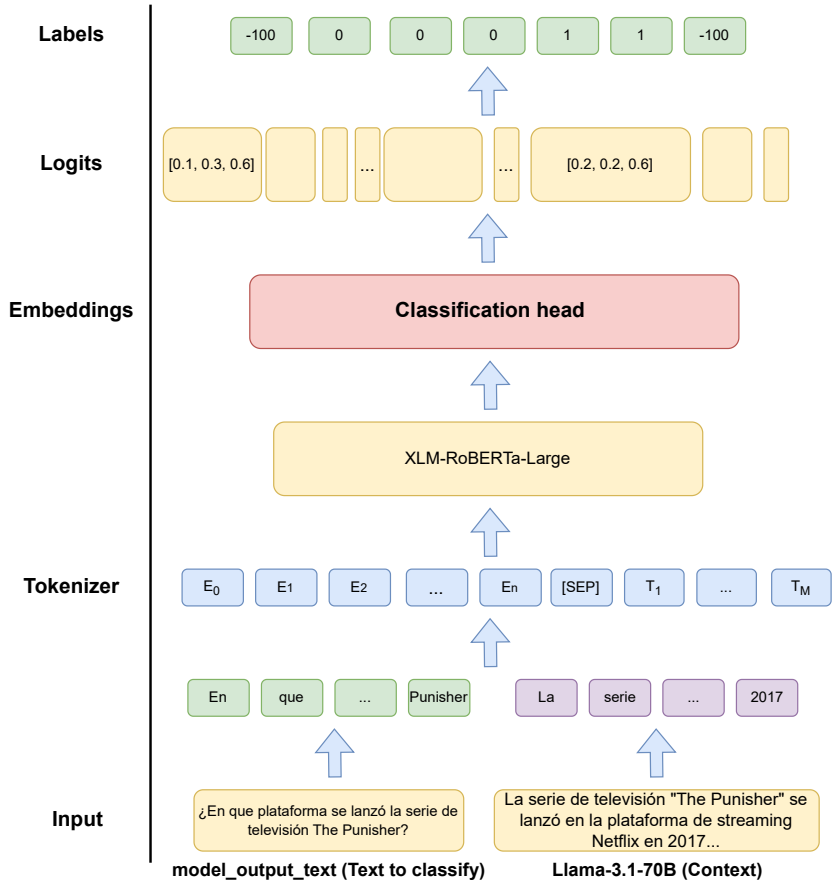

Figure 1: System architecture. Inputs are encoded using text_pair format: the first segment is the model output to classify (model_output_text), and the second is a hallucination-free reference generated by Llama-3.1-70B.

related to the tendency of the LLM to generate longer responses, which could be an indication that the model is generating hallucinations. Longer responses may be associated with a higher probability of producing unrelated or incorrect information, whereas the LLM teacher, being more oriented to provide precise and concise responses, has a more controlled response length.

For model training, the dataset has been divided into 80% for training and 20% for validation, in order to evaluate the model performance on an unseen dataset. As for the training parameters, a batch size of 16 tokens per device is used, with a total of 10 training epochs and a learning rate of 2e-5. Model evaluation is performed at the end of each epoch, and the model is saved after each evaluation.

To generate the answers with Llama-3.1-70B, the "meta-llama/Llama-3.1-70B-Instruct" template loaded in 4-bit format using the BitsAndBytes configuration was used to optimize memory usage on

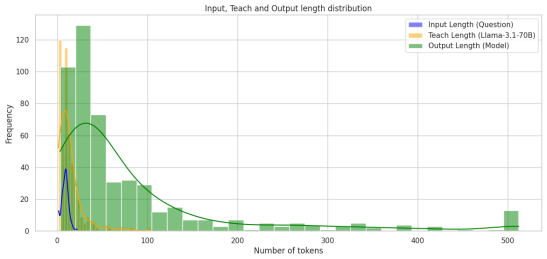

Figure 2: Input, Teach and Output length distribution.

GPUs.

The prompt used follows a consistent template specifying that the wizard must answer in the same language as the question, ensuring a clear, precise and concise answer. The prompt template is described in Listing 1.

For text generation, the following decoding parameters were used: a limit of 4096 tokens for the maximum length of the generated response, a top_p value of 0.95 to apply nucleus sampling and keep diversity controlled in the generation, and a top_k

```
You are an advanced multilingual assistant.
Your task is to answer the given
question in the same language as
the question. Ensure your response
is clear, accurate, short and concise.

Question: {question}

Answer:
```

Listing 1: Structure of the prompt

of 10 to restrict the selection to the 10 most likely tokens. In addition, a temperature of 0.7 was used to balance creativity and consistency of responses, along with a num_beams value of 1 to generate responses efficiently without beam search.

5 Results

Table 1 presents the official results of our approach compared to the baselines (mark all, mark none, and mark neutral) in terms of the IoU and Cor metrics. Performance was evaluated in fourteen languages, allowing for a comprehensive analysis of the effectiveness of our approach in a multilingual context.

Overall, our approach consistently outperforms the baselines in both metrics, indicating its ability to effectively detect LLM model-generated hallucinations in multiple languages. For example, in ES and EN, our model obtained an IoU of 0.2980 and 0.3667, respectively, and a Cor of 0.4152 and 0.4966. These results reflect a considerable improvement compared to the baselines, especially in the Cor metric, suggesting a higher accuracy in detecting token-level hallucinations.

It is observed that languages with more complex grammatical structures or less represented in the model pre-training, such as AR, FA and FI, show slightly lower performance. This behavior can be attributed to the lower availability of data in these languages, which affects the model's ability to generalize correctly. Moreover, only the "mark everything as hallucination" baseline outperforms our approach in some cases.

Notably, in languages such as SV and ZH, the model achieved an IoU of 0.4393 and 0.3875, respectively, with Cor values of 0.3936 and 0.4916. Although our approach presents a lower IoU than the baseline (mark all) in these languages, the model maintains a robust correlation, indicating its ability to capture consistent patterns in hallucination generation.

In terms of ranking, our approach demonstrates competitive performance, consistently ranking high for most of the languages evaluated. For example, it was ranked 13th in FA and 16th in IT, reflecting its effectiveness in languages with different linguistic backgrounds.

In conclusion, the results demonstrate the effectiveness and robustness of our approach for hallucination detection in multiple languages, significantly outperforming the baselines and maintaining a robust correlation. This confirms its generalizability in multilingual contexts and its potential application in language model-generated text evaluation tasks.

6 Conclusion

Our participation in Task 3 Mu-SHROOM of SemEval 2025 focused on the detection of hallucinations in texts generated by LLMs, a critical challenge as these systems are increasingly integrated into real-world applications. We adopted a token classification approach using the pre-trained multilingual XLM-RoBERTa-large model, augmented with contextual information from Llama-3.1-70B. By comparing the model-generated output with a hallucination-free reference, our system produces token-level hallucination probabilities, allowing for fine-grained analysis. Experimental results show that our approach consistently outperforms baseline methods in several languages, with particularly strong performance in English and Spanish. However, results in languages with complex grammar or fewer pre-training representations, such as Arabic and Finnish, highlight the need for language-specific strategies to improve generalization. These results highlight the benefits of combining multilingual modeling with LLM-generated context. Future work will explore more advanced fusion techniques, cultural and linguistic factors influencing hallucination generation, dataset expansion, and the use of reinforcement learning to further improve performance.

Future work will explore fine-tuning strategies such as cross-lingual transfer learning or data augmentation to mitigate the performance gap observed in low-resource languages. In addition, while Llama-3.1-70B provides strong contextual guidance, its size may hinder real-time deployment. Exploring more lightweight alternatives, such as distilled models or hybrid architectures, remains a promising direction. Although we did not perform

Table 1: Official results for each language (L), reporting the rank (#), Cor of baseline (mark all, mark none, and mark neutral), IoU and Cor.

L	#	IoU Bas.	IoU Bas.	IoU Bas.	IoU	Cor
		mark all	mark none	mark neutral		
AR	20	**0.3614**	0.0467	0.0418	0.3436	0.4211
CA	16	0.2423	0.0800	0.0524	**0.4301**	0.4295
CS	16	0.2632	0.1300	0.0957	**0.3380**	0.3600
DE	18	0.3451	0.0318	0.0267	**0.4093**	0.4403
EN	26	0.3489	0.0325	0.0310	**0.3667**	0.4966
ES	16	0.1853	0.0855	0.0724	**0.2980**	0.4152
EU	20	**0.3671**	0.0208	0.0101	0.3272	0.3925
FA	13	0.2028	0.0000	0.0001	**0.4677**	0.3939
FI	20	**0.4857**	0.0000	0.0042	0.4563	0.5126
FR	26	**0.4543**	0.0000	0.0022	0.3200	0.4117
HI	17	0.2711	0.0000	0.0029	**0.4510**	0.4386
IT	16	0.2826	0.0000	0.0104	**0.4413**	0.4601
SV	17	**0.5373**	0.0204	0.0308	0.4393	0.3936
ZH	17	**0.4772**	0.0200	0.0236	0.3875	0.4916

an ablation study to isolate the specific contribution of contextual input, this is an important avenue for future work to better understand its impact on hallucination detection.

Acknowledgments

This work is part of the research project LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MCIN/AEI/10.13039/501100011033 and the European Fund for Regional Development (ERDF)-a way to make Europe, and the research project "Services based on language technologies for political microtargeting" (22252/PDC/23) funded by the Autonomous Community of the Region of Murcia through the Regional Support Program for the Transfer and Valorization of Knowledge and Scientific Entrepreneurship of the Seneca Foundation, Science and Technology Agency of the Region of Murcia. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 912–928.

Fadi Aljamaan, Mohamad-Hani Temsah, Ibraheem Altamimi, Ayman Al-Eyadhy, Amr Jamal, Khalid Alhasan, Tamer A Mesallam, Mohamed Farahat, Khalid H Malki, et al. 2024. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Medical Informatics*, 12(1):e54345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2025. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578.

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

José Antonio García-Díaz, Ronghao Pan, and Rafael Valencia-García. 2023. Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in spanish and english. *Mathematics*, 11(24).

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2024. Zero-resource hallucination prevention for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3586–3602.

Timothee Mickus, Elaine Zosa, Raúl Vázquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. Semeval-2024 shared task 6: Shroom, a shared-task on hallucinations and related observable overgeneration mistakes. *arXiv preprint arXiv:2403.07726*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2024. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. *CMES-Computer Modeling in Engineering & Sciences*, 140(3).

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2025. Spanish mtlhatecorpus 2023: Multi-task learning for hate speech detection to identify speech type, target, target group and intensity. *Computer Standards & Interfaces*, 94:103990.

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.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

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.

Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and Jiliang Tang. 2024. Ai for education (ai4edu): Advancing personalized education with llm and adaptive learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6743–6744.

Ma Xiaoliang, Zhao RuQiang, Liu Ying, Deng Congjian, and Du Dequan. 2024. Design of a large language model for improving customer service in telecom operators. *Electronics Letters*, 60(10):e13218.