YNU-HPCC at SemEval-2025 Task3: Leveraging Zero-Shot Learning for Hallucination Detection

Shen Chen, Jin Wang, and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
chenshen@stu.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

Abstract

This study reports the YNU-HPCC team's participation in SemEval-2025 shared task 3, which focuses on detecting hallucination spans in multilingual instruction-tuned Large Language Models (LLMs) outputs. This task differs from typical hallucination detection tasks in that it does not require identifying the entire response or pinpointing which sentences contain hallucinations generated by the LLM. Instead, the task focuses on detecting hallucinations at the character level. In addition, this task differs from typical hallucination detection based on binary classification. It requires not only identifying hallucinations but also assigning a likelihood score to indicate how likely each part of the model output is hallucinatory. Our approach combines Retrieval-Augmented Generation (RAG) and zero-shot methods, guiding LLMs to detect and extract hallucination spans using external knowledge. The proposed system achieved first place in Chinese and fifteenth place in English for track 3[1].

1 Introduction

Hallucination in large language models refers generating of information that appears plausible but is factually incorrect or fabricated. This issue is common in open-domain tasks, such as question answering and summarization, where the model may produce answers inconsistents with the provided context or external knowledge.

Hallucinations can be categorized into two main types (Ji et al., 2023): intrinsic, which conflicts with the source content, and extrinsic, which cannot be verified from the source content. These errors are closely related to the nature of knowledge. Some knowledge is static, such as the date of the Civil War, while other knowledge evolves over time, such as the current population of China. The distinction between these two types of knowledge implies that hallucinations cannot be eliminated—especially for the latter—unless the model adopts a Retrieval-Augmented Generation (RAG)-based (Lewis et al., 2020) approach.

As a result, much effort has been dedicated to hallucination detection. These detection methods can be broadly classified into the following categories (Huang et al., 2025): 1) Methods based on model output logits, such as uncertainty and semantic entropy; 2) Fact-based detection methods, which generally calculate the similarity between factual documents and the model's output. The shortcomings of both approaches are evident: the former can only detect hallucinations but cannot correct them, and it requires the detection process to be tightly coupled with model generation. The latter, on the other hand, struggles in domains where factual documents are difficult to retrieve.

This task evaluated whether the model's output addressed the question and whether the answer was accurate. Neither of these approaches effectively addresses the task's minimum interval requirement because both focus on the entire model output rather than specific spans within the answer.

Therefore, our team's approach initially combined fact-based documentation with Machine Reading Comprehension (MRC) (Kenton and Toutanova, 2019), followed by integrating the factual document-based method with zero-shot detection. The final experimental results demonstrated that our solution was both effective and competitive (Vázquez et al., 2025).

2 Related Work

Machine Reading Comprehension. Machine reading comprehension is often applied in tasks that answer questions based on context, and we similarly use it to detect erroneous content. As men-

[1]Our code is available at https://github.com/deepdarklowtech/YNU-HPCC-SemEval2025-Task3

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 28–33

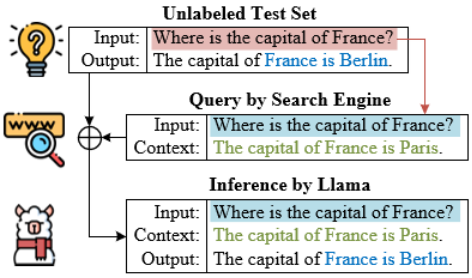

Figure 1: Combining MRC and RAG system architecture approach

tioned in the Introduction, the task requires identifying hallucinations at the character-level interval, rather than evaluating the entire answer, which is more common in hallucination detection.

Zero-shot Learning. (Brown et al., 2020) The Zero-Shot Prompting approach is highly flexible and generalizable, eliminating the need for task-specific training across new tasks or domains. Instead, it relies on pre-trained language models combined with carefully designed examples or prompts to facilitate reasoning and generate outputs. Although the validation set for the task initially contained up to 807 Q&A pairs when categorized by language, filtering for unique questions reduced the dataset to around 50 distinct entries. Given this constraint, we adopted the zero-shot approach as our solution.

3 Approach

3.1 MRC Combined with RAG for Hallucination Detection

Our initial approach combined MRC techniques with an RAG strategy to label hallucination intervals,as shown in the flow in Figure1. The core of this approach focused on the retrieval-augmented component. While interfaces like Google and Bing required extensive data cleansing, our team opted for a more direct method: we manually queried the model input field. We filtered the generated answers to ensure they were concise and directly addressed the question. Additionally, since the interval-based hallucination detection dataset is only available in English, the fine-tuned LLM must possess cross-linguistic capabilities. For this reason, we selected LLaMA3 (Dubey et al., 2024) as the model for our approach.

To meet the task's minimum interval requirement, we used BIO tagging (Ramshaw and Marcus, 1999) for individual tokens. However, a limitation of this approach was that it did not allow us to populate the soft labels field with probability values. To overcome this, we used the softmax value of the hallucination end token as a proxy. This decision was informed by two factors: (1) LlamaForTokenClassification[2], uses Cross-Entropy as the loss function for multi-label training, and (2) marking the end token marginally improved the CoR score in the final evaluation.

3.2 Zero-shot Combined with RAG for Hallucination Detection

Following the previous phase, we created a question-answer pair document based on the test set. In the next step, we applied prompt engineering to guide the LLM in directly categorizing the contents of the model output text field. The model was instructed to output the soft and hard labels fields separately, as shown in the flow in Figure2. Our team selected four models for this solution: OpenAI-o1, Claude-3.5, Gemini and DeepSeek V3 (Liu et al., 2024).

We also explored using the task-provided training set, along with data from the validation set or open-source datasets (e.g., RAGTruth (Niu et al., 2023) and HaluEval (Li et al., 2023)) for a few-shot learning approach. However, based on our previous experience with MRC, we found that some entries in the official dataset contain hallucination spans shorter than a token. Additionally, open-source datasets often lack the probability values required for the soft labels field. While we attempted to annotate these datasets with probability values, we encountered a challenge: assigning error probabilities to clearly incorrect content and assigning probabilities to valid content. Furthermore, when the text is known to be incorrect, it is challenging for the LLM to provide the required probability values for the soft labels field.

The task organizers explicitly stated that 12 reviewers annotated the probability values for both English and Chinese. Therefore, using a prompt, we instructed the model to select a probability value between 0.0833 (1/12) and 1.0 (12/12) for the soft labels field. The organizers' statement aligns with the challenges we described in the previous paragraph. This method of assigning probabilities is

[2] https://huggingface.co/docs/transformers/en/model_doc/llama#transformers.LlamaForTokenClassification

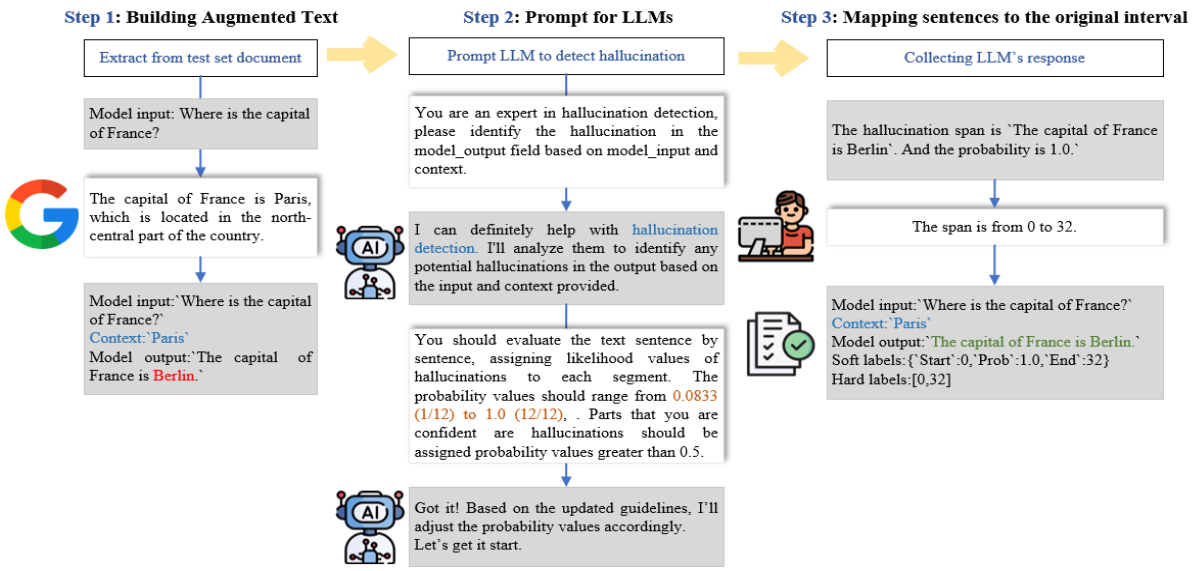

Figure 2: Combining zero-shot and RAG system architecture approach

virtually infeasible, as it relies on aggregating multiple binary labels to approximate probability values.

Both models naturally use sentences as the unit for the minimum hallucination span during the response collection process. Although this approach does not fully meet the "minimum" span requirement set by the organizers, it aligns with our team's intuition about hallucinations. We typically don't distinguish between correct and incorrect information at the word level; instead, we usually make distinctions at least at the sentence or clause level.

As widely known, the strawberry challenge revealed LLMs' weakness in counting. The results are often unreliable when directly asked to generate soft and hard labels in JSON format. Fortunately, both models apply chain-of-thought (CoT) (Wei et al., 2022) reasoning to generate responses, improving accuracy. To mitigate counting errors in the LLM outputs, we required both OpenAI-o1 and DeepSeek V3 to provide the spans and the corresponding textual content for each span. After collecting the LLM responses, we used the KMP algorithm to map the text to its corresponding position in the model output text field.

4 Experiment Detail

Datasets. In our MRC solution, we fine-tuned the LLaMA model using RAGTruth and HaluEVAL datasets. RAGTruth explicitly labels hallucination intervals in numerical form, closely aligning with the task requirements, while HaluEVAL only identifies the text segments containing hallucinations without specifying precise intervals.

Evaluation Metric. This task adopts IoU and Cor as evaluation metrics during the assessment phase. The scoring criteria for IoU are outlined below:

$$IoU = \frac{\text{Intersection}(Gold, Prediction)}{\text{Union}(Gold, Prediction)} \quad (1)$$

where *Gold* refers to the intervals marked by the organizers as hard labels, while *Prediction* refers to the intervals marked by participants as hard labels. Cor's score was calculated using the Spearman Rank Correlation Coefficient, which is calculated as follows:

$$Cor = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (2)$$

where d_i refers to the difference in probability values between the soft labels published by the organizers and the soft labels submitted by the participants, while n represents the character-level length of all soft label intervals marked by the organizers.

5 Result

In the final evaluation phase, the tournament organizers used Intersection-over-Union (IoU) to evaluate the hard_labels, while Spearman correlation was used to assess the soft_labels.Table 2 and 3 presents the results achieved using different models and methods.

As shown in the table 1 and 2, OpenAI-o1 outperforms the other models by a significant margin.

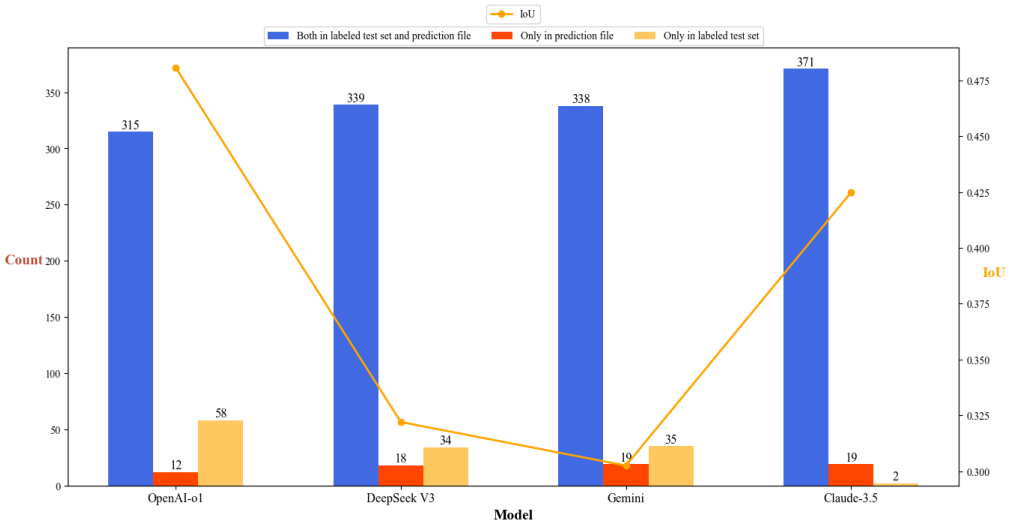

Figure 3: Performance of different LLMs on the EN test set. Blue bars represent sentences correctly labeled by the model, red bars indicate sentences incorrectly labeled, and yellow bars show sentences that the model failed to detect.

LLM	IoU	Cor
OpenAI-o1(Prompt)	**0.5540**	**0.3518**
DeepSeek V3(Prompt)	0.1219	0.0497
Gemini V2(Prompt)	0.3265	0.0664
Claude-3.5(Prompt)	0.4769	0.0795
LLaMA-3(MRC)	0.4565	0.1846
Baseline(mark all)	0.4772	0.0000

Table 1: Results obtained with different LLMs in the ZH test set.

LLM	IoU	Cor
OpenAI-O1(Prompt)	**0.4807**	**0.4075**
DeepSeek V3(Prompt)	0.3220	0.1802
Gemini V2(Prompt)	0.3025	0.1722
Cluade-3.5 (Prompt)	0.4248	0.3391
LLaMA-3(MRC)	0.3800	0.3974
Baseline(mark all)	0.3489	0.0000

Table 2: Results obtained with different LLMs in the EN test set.

The MRC-based approach using LLaMA for hallucination interval detection also performs well.

To better demonstrate the effectiveness of our team's solution, we also performed a more refined data analysis at the end of the evaluation phase. Our analysis is based on the labeled test set released by the organizers at the end of the evaluation phase. As we stated in Section 3.2, the granularity of the labels provided by the organizers is lower than our

Model	Rank	IoU	Cor
Chinese(Mandarin)			
OpenAI-o1	1	0.5540	0.3518
Claude-3.5	5	0.4769	0.0795
LLaMA-3	14	0.4565	0.1846
Gemini V2	21	0.3265	0.0664
DeepSeek V3	26	0.1219	0.0497
English			
OpenAI-o1	15	0.4807	0.4075
Cluade-3.5	24	0.4248	0.3391
LLaMA-3	26	0.3800	0.3974
DeepSeek V3	32	0.3220	0.1802
Gemini V2	35	0.3025	0.1722

Table 3: Ranking of our practices in the official ranking table

team's judgment of the LLMs' hallucination phenomenon, and also Tables 3 have demonstrated the accuracy and relevance at the character level. Therefore, we split the model output text by sentence and analyze it at the sentence level, as shown in Figure 3 and Figure 4. In terms of detection ability at the sentence level, Claude-3.5 performs ahead of all other models. However, OpenAI-o1 scores higher under the IoU metric due to its fewer false positive results. Meanwhile, Figures 3 and 4 show a disproportionate increase in sentences that were not detected by the Gemini and DeepSeek V3 models in the Chinese track. However, the opposite trend is observed under the IoU metric. This dis-

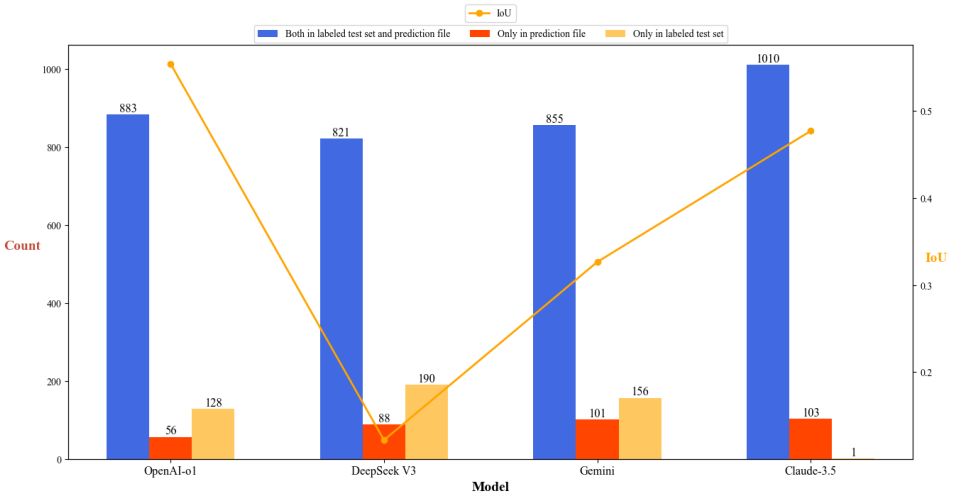

Figure 4: Performance of different LLMs on the ZH test set. Blue bars represent sentences correctly labeled by the model, red bars indicate sentences incorrectly labeled, and yellow bars show sentences that the model failed to detect.

crepancy is primarily due to the fact that the model responses in the Chinese test set often include more Markdown syntax for structured, point-by-point responses to questions. This is also reflected in the overall length of the text, which is 48040 for the Chinese test set and 36745 for the English dataset. As a result, regular expressions struggle to segment the text according to human linguistic conventions.

6 Analysis

The data in the table indicates a low correlation in our solution, which can be attributed to at least the following factors:

- Following the example provided by the task organizers, if the text in the model output text is *The capital of France is Berlin*, the hallucination interval we provide should only include the token "Berlin." This tokenization approach is, of course, correct. We replace *Berlin* with *Paris* to correct the LLM's response.

- Similarly, for hallucination detection, fitting the probability of a binary classification task to the results derived from 12 individuals' votes proves too challenging. We also considered using multiple rounds of sampling (Shanahan et al., 2023), where the discriminative results of 12 judgments made by the same model would be used to fit the final probability; however, this exceeded our team's budget.

- As shown in Tables 1 and 2, our team's results in the English track were not as strong as those in the Chinese track, which seems contradictory considering the training data used by the relevant models. This discrepancy may be related to the quality of the augmented documents we created. Specifically, for the test set, there were 12 instances in the English portion where content could not be retrieved or was overly verbose, while only 4 were present in the Chinese portion. This difference could have a significant impact, given that the test set contains only 150 entries.

7 Conclusion

This study describes the work conducted by the YNU-HPCC team for participation in "MuSHROOM (SemEval 2025)." The methods we employed include Augmented Document-based MRC and Augmented Document-based Prompt Engineering. The final results showed that using OpenAI-o1 with Augmented Document-based Prompt Engineering achieved first place in the Chinese track, with an IoU score of 0.5540 and a Cor score of 0.3518. In the English track, the model achieved 15th place, with an IoU score of 0.4807 and a Cor score of 0.4075.

Future work attempts to adopt a knowledge graph-based approach for self-checking LLM-generated answers in the future. After all, compared to a simplistic model, a more intelligent model that outputs incorrect answers tends to cause more significant harm to society and the economy.

Therefore, integrating online retrieval-augmented generation (RAG) with offline knowledge graphs will be key for mitigating hallucination.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051. We would like to thank the anonymous reviewers for their constructive comments.

References

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.

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*.

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

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

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.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396*.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.