NLP_CIMAT at SemEval-2025 Task 3: Just Ask GPT or Look Inside. A prompt and Neural Networks Approach to Hallucination Detection

Jaime Stack-Sánchez, Miguel A Alvarez-Carmona and Adrián Pastor López-Monroy
Mathematics Research Center (CIMAT)
jaime.stack@cimat.mx
miguel.alvarez@cimat.mx
pastor.lopez@cimat.mx

Abstract

This paper presents NLP_CIMAT's participation in SemEval-2025 Task 3 (Vázquez et al., 2025), which focuses on hallucination detection in large language models (LLMs) at character level across multiple languages. Hallucinations—outputs that are coherent and well-formed but contain inaccurate or fabricated information—pose significant challenges in real-world NLP applications. We explore two primary approaches: (1) a prompt-based method that leverages LLMs' own reasoning capabilities and knowledge, with and without external knowledge through a (RAG)-like framework, and (2) a neural network approach that utilizes the hidden states of a LLM to predict hallucinated tokens. We analyze various factors in the neural approach, such as multilingual training, informing about the language, and hidden state selection. Our findings highlight that incorporating external information, like wikipedia articles, improves hallucination detection, particularly for smaller LLMs. Moreover, our best prompt-based technique secured second place in the Spanish category, demonstrating the effectiveness of in-context learning for this task.

1 Introduction

Since the introduction of the transformer architecture in 2017 (Vaswani et al., 2017), large language models have rapidly advanced, finding applications in both scientific research and everyday life. However, these models face two major challenges (Mickus et al., 2024): They often generate false or misleading information that appears syntactically correct and current evaluation metrics prioritize fluency and grammatical accuracy over factual correctness.

This combination leads to what is known as hallucination, that is, where models produce outputs that are coherent and well-formed but contain inaccurate or fabricated information—an issue that remains difficult to detect automatically. Hallucinations pose a significant barrier to the practical development of LLMs and their mass adoption as reliable tools in everyday life.

Although in this work we treat hallucination detection as a task, hallucinations can appear in various domains and some works address hallucination detection in fields like machine translation (Dale et al., 2022; Guerreiro et al., 2023), summarization (Huang et al., 2021; Van der Poel et al., 2022), definition modeling (Mickus et al., 2024) and dialogue generation (Lei et al., 2023), we are still far from establishing a unified dataset and system for hallucination detection.

Despite some progress in hallucination detection, current works and datasets do not attack the problem with a granularity that allows one to know exactly where the hallucination is in the text. SemEval-2025 task 3 introduces a test bed that allows us to tackle the problem with character level granularity and information to compare the correlation between the proposed systems and human annotations. The task is closely related to fact checking and consist of identifying the parts of the model output that are hallucinated at character level given the model input. The dataset includes 14 languages where every instance indicates the language, the ranges of hallucinated characters and the probability assigned to these hallucinations.

This paper presents the participation of NLP_CIMAT in the shared task which consist on the development and study of two main approaches. The first one is a prompt-based method that leverages the intrinsic knowledge and capabilities of the LLM, where the model is directly asked to highlight the hallucinated parts of its output. We explore two key variants of this method: Without external knowledge – The model relies solely on its internal knowledge; and with external knowledge (RAG-like framework) – We investigate whether incorporating retrieved external information im-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1683–1689

proves hallucination detection performance.

The second approach is a neural network that leverages the encoded information in the internal hidden state representations of a LLM to predict whether a token is hallucinated. This method explores several key aspects to optimize performance:

1. Individual vs. Multilingual Training – We investigate whether training on a single language or across multiple languages leads to better generalization.

2. Incorporating a Language Vector – We assess whether adding a one-hot encoded vector indicating the language improves model performance.

3. Number of Parameters – We analyze whether a larger classifier architecture (more layers and neurons) leads to better hallucination detection.

4. Hidden State Layer – We determine which hidden state layer contains the most relevant information for hallucination prediction.

5. Concatenating Multiple Hidden States – We evaluate whether using hidden states from multiple layers enhances model performance compared to using a single layer.

2 Related Works

Prompt-based techniques represent the state of the art in hallucination detection, with most top-performing approaches in the SemEval 2024 Shared Task 6 (Mickus et al., 2024) relying on prompt-based methodologies to achieve strong results.

SELFCHECKGPT (Manakul et al., 2023): utilizes a sample based strategy to generate multiple stochastic samples. This work proposes that a model with a good understanding and knowledge of the task or concept is less likely to generate inconsistent information and hallucinations. This work demonstrates the effectiveness of prompt based approaches to detect hallucinations, we took inspiration in their prompt based approach modifying their structure to not rely on samples.

Fact-checking performance of LLMs improves notably when they are given contextual information, as shown by Quelle and Bovet (2024); Krishnamurthy and Balaji (2024). LLMs can effectively leverage external knowledge to generate responses and support claims with factual accuracy. We took inspiration from their work and provided our prompt approach with external information retrieved from wikipedia.

In recent years there have been numerous studies about using hidden states of a LLM to detect hallucinations. Azaria and Mitchell (2023) train a MLP on the hidden states of a LLM to detect hallucinations at sentence level and investigate which hidden states contain relevant information to correctly classify hallucinations. Similarly, Duan et al. (2024) analyze the changes in the internal states of a LLM when it generates factual versus non-factual claims, using these differences to determine whether a hallucination has occurred. Our work draws inspiration from both studies: we adopt the MLP architecture from the first and leverage insights from the second to develop a token-level hallucination classifier.

3 Methodology

In this section we will introduce our proposed systems for hallucination detection, dividing them in two groups: Prompt based approach and Hidden States Neural Network approach.

3.1 Prompt based approach

The core idea behind these methods is to leverage the inherent knowledge and reasoning capabilities of LLMs to detect hallucinations effectively.

We present two prompt based approaches for hallucination detection:

- Few-shot without external knowledge – This method relies solely on the intrinsic capabilities and knowledge of the model to classify hallucinated characters in a response.

- Few-shot with external knowledge – In this approach, Wikipedia articles are retrieved as external knowledge, allowing the model to combine its internal knowledge with up-to-date factual information to improve classification accuracy.

In all of our submissions for the prompt based approach we used a few shot scheme to reduce the probability of the LLM generating an answer that we couldn't analyze automatically. The models we used were gpt-4o and gpt-3.5-turbo.

We decided to use these models because there have been multiple studies showing their great capabilities on solving a diverse amount of tasks (Chen et al., 2024) and they are trained in recent data,

"You are an expert fact checker, your job is to answer with the parts of the hypothesis that contain inaccuracies or false information given the context provided and the query. Answer with the exact same text as the hypothesis but using < > to highlight the incorrect parts of the hypothesis.
For example:
Context: [wikipedia page]
Query: [Model Input]
Hypothesis: [Model output]
Answer: [Model output with hallucination highlighted]

Context: [wikipedia page]
Query: [Model Input]
Hypothesis: [Model output]
Answer: [Model output with hallucination highlighted]"

Figure 1: Prompt scheme: Highlighted in green we have the Role, in red we have the Examples (3) and last in purple the Context (optional)

which makes them fit for the task. We also aim to compare the performance of these models to assess the impact of model size and the quality of training data on hallucination detection.

3.1.1 Few-shot without external information

The idea is to give the model an instance composed on the original input and output with a few examples and ask it to directly, without any additional information, tell us where the hallucinations are.

For the prompt construction we first gave a role for the model, which has been shown to improve the models capabilities compared to when no role is given. Then we followed with the format in which the answer was to be given, we opted to indicate the model to answer with the same exact text as the model output but highlighting the hallucinated parts of the output. Then we gave three examples to the model, tackling the three different possible cases: the output has one hallucination, the output contains no hallucinations and the output contains multiple hallucinations. The prompt scheme can be seen in Figure 1.

3.1.2 Few-shot using external information

RAG consists in providing the model with external information that can help to answer the task that the model is given. We hypothesize that giving the model extracts from wikipedia can improve the model performance to identify the incorrect information from the model output. To extract the wikipedia page we used the Model input followed by the word "Wikipedia" and then we proceeded to retrieve the first wikipedia link we found in google search, later retrieving the unformatted wikipedia text and gave it, completely, to the model as context.

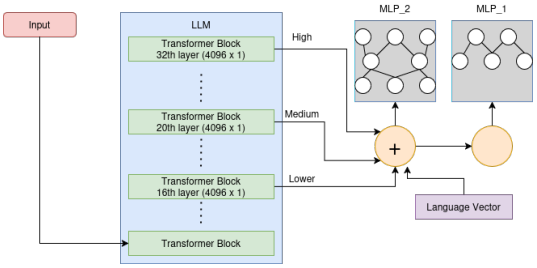

Figure 2: Framework for the Neural Network approach. We extract hidden state vectors from three different transformer blocks, concatenate them, and add a Language Vector (LV). The combined representation is then fed into our MLP classifiers for hallucination detection.

3.2 Hidden States Neural Network approach

The proposed model takes the encoded information on the extracted hidden states of the LLM that contains relevant information to correctly classify an hallucination. To extract the hidden states, we structured the sequence that we pass to our LLM like: "[Model input] [Model output]".

We focused on developing an MLP model trained on hidden states from LLaMA 3.1 8B Instruct. We extract and concatenate up to three hidden states, (H, M, L)[1], incorporating a Language Vector (LV)—a one-hot encoded vector of length 10, representing the 10 languages in the validation dataset. This vector was concatenated at the beginning of the hidden states. The resulting representation was then passed to the MLP model, which predicted whether the corresponding token was hallucinated. Our framework is illustrated in figure 2.[2]

We experimented with different training configurations, including:

- Training the model on all languages (Multilingual) vs. individual languages (M or I)
- Using one hidden state vs. concatenating three hidden states
- Adding or omitting the Language Vector (LV)

[1] 32, 20 and 16

[2] For all the models we trained with a batch size of 16, a learning rate of 1e-4 and cross entropy loss as the loss function.

4 Results

We first present the best results submitted for each language, followed by a study analyzing the impact of the different strategies we implemented.

4.1 Best prompt results

For the prompt-based approach, we focused only on Spanish and English due to time and budget constraints. The objective of this experiment is to evaluate the effectiveness of prompt-based approaches for hallucination detection in LLMs. Specifically, we analyze the impact of external knowledge integration and compare the performance of models of different sizes. Table 1 presents our best prompt results, with one of our Spanish submission achieving second place.

Language	IoU	Cor	RAG	Model
***ES**	**0.520**	**0.523**	**TRUE**	**gpt-4o**
ES	0.518	0.520	FALSE	gpt-4o
ES	0.353	0.351	TRUE	gpt-3.5-turbo
ES	0.267	0.253	FALSE	gpt-3.5-turbo
EN	0.457	0.370	TRUE	gpt-4o
EN	0.434	0.415	FALSE	gpt-4o
EN	0.328	0.341	TRUE	gpt-3.5-turbo
EN	0.299	0.291	FALSE	gpt-3.5-turbo

Table 1: Best prompt results. Incorporating external knowledge (RAG) consistently improved performance, especially for GPT-3.5-Turbo, with an 8.6% IoU gain. GPT-4o achieved overall better results. *Second place winner in the spanish category

From Table 1, we observe that RAG has a greater impact on GPT-3.5-Turbo than on GPT-4o. In GPT-3.5-Turbo, the IoU gain is more significant, reaching 8.6% higher, whereas in GPT-4o, the maximum difference is only 2.3%.

Additionally, as expected, GPT-4o significantly outperforms GPT-3.5-Turbo in hallucination detection. Interestingly, GPT-3.5-Turbo with RAG still couldn't surpass the results of GPT-4o without RAG, suggesting that GPT-4o's superior architecture and training enable better information retrieval and utilization, even without external augmentation.

4.2 Best Hidden States Neural Network results

The objective of this experiment is to evaluate effectiveness of using hidden states from a LLM to detect hallucinations at token level. Table 2 presents our best submitted results for the Hidden States Neural Network Approach, selected from a set of experiments with varying parameters. In the Concat Layers column, the letters H, M, and L represent the hidden states extracted from layers 32, 20, and 16, respectively.

From the table we can see that multilingual appeared much more in the best results, giving us an idea that training with data in all languages can improve the performance of the models. Concatenating layers doesn't appear to have a significant impact in the results, and we can observe that using more layers in our MLP appears to yield better performance. But we will explore this ideas in the following sections.

Language	IoU	Cor	IoU Bas. mark all	Cor Bas. mark all	M or I	Layers	# Layers	LV	Epoch
Arabic	0.204	0.077	0.316	0.007	I	[H,L,M]	2	False	5
Catalan	0.141	0.069	0.242	0.06	M	L	2	False	5
Chinese	0.220	0.145	0.477	0	M	L	3	False	5
Basque	0.175	0.052	0.367	0	M	L	3	False	5
Farsi	0.0316	0.394	0.203	0.01	M	L	2	False	15
Finnish	0.374	0.031	0.486	0	M	[H,L,M]	3	True	15
French	0.353	0.071	0.454	0	M	[H,L,M]	3	True	15
Italian	0.189	0.045	0.283	0	I	[H,L,M]	3	False	15
Swedish	0.238	0.054	0.537	0.014	I	[H,L,M]	2	True	15
English	0.174	0.129	0.349	0	M	[H,L,M]	3	TRUE	15
Spanish	0.111	0.092	0.185	0.013	M	L	3	TRUE	15

Table 2: Best results Hidden States Neural Network approach

4.3 Study of internal structure of the MLP models

The objective of this experiment is to analyze the impact of the complexity of the MLP along with which hidden state configuration provides the most useful information for the neural network to make accurate classifications. Tables 3,4 and 5 present the results of varying internal parameters of our classifier. We focused on two sets of languages: Languages with abundant training data in our model and languages with limited training data in our model.

For the well represented languages we chose English and Spanish, for the languages with limited training data we chose Finnish and Swedish. This distinction allows us to analyze in a more meaningful way the impact of varying the internal structure of our classifier.

We observe clear differences in the achieved metrics. While our preliminary results on the validation dataset suggested that layer L was the most effective, the test dataset results indicate that layer H performed best, achieving the highest overall metrics. A possible explanation for this is that layer H, being the last layer, encodes information closely related to the logits of the generated token. These logits reflect the probability distribution assigned by the model to the generated token. If the model is uncertain about its response, this probability distribution is likely to shift compared to a more confidently generated token. The classifier can leverage these probability variations to improve the identification of hallucinated tokens.

Surprisingly, concatenating multiple layers does not improve results compared to using only layer H. This outcome was unexpected. One possible explanation is that the MLP model is relatively small and may not have the capacity to effectively utilize the additional information provided by concatenating three hidden states. For this same reason it is not surprising that the model with 3 hidden layers outperformed the model with 2 hidden layers.

Language	IoU	Cor	Layers
English	0.142	0.146	L
English	0.151	0.140	M
English	0.173	0.151	H
Spanish	0.116	0.093	L
Spanish	0.088	0.097	M
Spanish	0.086	0.058	H
Finnish	0.311	0.036	L
Finnish	0.339	0.026	M
Finnish	0.330	0.032	H
Swedish	0.133	0.062	L
Swedish	0.170	0.078	M
Swedish	0.191	0.064	H
All languages	0.172	0.120	L
All languages	0.183	0.114	M
All languages	0.190	0.084	H

Table 3: Results of the MLP considering: One hidden state, multilingual with LV, 2 hidden layers in MLP and trained for 15 epoch

Language	IoU	Cor	Layers
English	0.133	0.145	L
English	0.155	0.139	M
English	0.189	0.114	H
Spanish	0.112	0.092	L
Spanish	0.105	0.077	M
Spanish	0.095	0.050	H
Finnish	0.268	0.035	L
Finnish	0.330	0.025	M
Finnish	0.364	0.040	H
Swedish	0.167	0.076	L
Swedish	0.166	0.069	M
Swedish	0.265	0.067	H
All languages	0.169	0.122	L
All languages	0.179	0.112	M
All languages	0.201	0.084	H

Table 4: Results of the MLP with: One hidden state, multilingual with LV, 3 hidden layers in MLP and trained for 15 epoch

4.4 Study of the relevance of information in the MLP models

In this experiment we analyze the impact of multilingual training and the addition of the language vector. In Table 6, we observe that while the best results were achieved using multilingual training with a language vector, the comparison between training on an individual language versus multilingual training without a language vector is less conclusive. The results do not show a clear advantage for either approach, suggesting that the impact of multilingual training without explicit language

Language	IoU	Cor	# Layers
English	0.158	0.135	2
English	0.175	0.130	3
Spanish	0.089	0.079	2
Spanish	0.092	0.070	3
Finnish	0.315	0.035	2
Finnish	0.374	0.031	3
Swedish	0.172	0.075	2
Swedish	0.179	0.082	3
All languages	0.171	0.106	2
All languages	0.190	0.108	3

Table 5: Results of the MLP with: Concatenated hidden states [H, L, M], multilingual with LV and trained for 15 epoch

information varies depending on the specific conditions.

We observe that for languages with abundant training data in our proxy model, multilingual training slightly improves performance. However, for languages with limited training data, training on multiple languages reduces performance.

We hypothesize that this occurs because, in well-represented languages in our model, the encoded information in the hidden states is more distinct and easily differentiable. In contrast, for languages with less training data, the encoded information is more subtle, making it harder for the model to generalize effectively across multiple languages.

Language	IoU	Cor	M or I	LV
English	0.118	0.089	I	Doesn't apply
English	0.128	0.105	M	False
English	0.189	0.114	M	True
Spanish	0.082	0.054	I	Doesn't apply
Spanish	0.093	0.037	M	False
Spanish	0.095	0.050	M	True
Finnish	0.315	0.035	I	Doesn't apply
Finnish	0.294	0.032	M	False
Finnish	0.364	0.040	M	True
Swedish	0.244	0.065	I	Doesn't apply
Swedish	0.214	0.058	M	False
Swedish	0.265	0.067	M	True
All languages	0.186	0.078	I	Doesn't apply
All languages	0.184	0.078	M	False
All languages	0.201	0.084	M	True

Table 6: Results of the MLP with: H layer, 3 hidden layers in MLP and 15 epoch of training

5 Conclusion

This work summarizes the approach of the NLP_CIMAT team in the SemEval 2025 Shared Task 3. We present two primary methodologies for hallucination detection: A prompt-based approach that leverages the capabilities and knowledge of LLMs to accurately identify hallucinated characters in a generated output; and a neural network approach trained on hidden states to classify whether a token is hallucinated.

Our findings indicate that incorporating external information in the prompt-based approach significantly improves performance for GPT-3.5-Turbo, while for GPT-4o the difference is negligible. We hypothesize that this discrepancy is closely related to model size and the quantity and quality of training data used. For the neural network approach, we identified the optimal configuration as follows: Using a three layer MLP architecture trained using the last hidden state layer and multiple languages while informing the classifier about the language. While the prompt-based method achieved better results, the model sizes used in the two approaches are not directly comparable.

A key direction for future work is to evaluate both methods using the same language model to determine which approach yields superior performance under equal conditions.

6 Acknowledgments

The authors thank Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), in particular Jaime Stack-Sánchez thanks SECIHTI for the scholarship becas nacionales para estudios de posgrado (#001738) (CVU: 1286277), Centro de Investigación en Matemáticas (CIMAT) and CIMAT Bajio Supercomputing Laboratory (#300832) for the computational resources provided.

References

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

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How is chatgpt's behavior changing over time? *Harvard Data Science Review*, 6(2).

David Dale, Elena Voita, Loïc Barrault, and Marta R Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal work-

ings alone do well, sentence similarity even better. *arXiv preprint arXiv:2212.08597*.

Hanyu Duan, Yi Yang, and Kar Yan Tam. 2024. Do llms know about hallucination? an empirical investigation of llm's hidden states. *arXiv preprint arXiv:2402.09733*.

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.

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*.

Vallidevi Krishnamurthy and Varshini Balaji. 2024. Yours truly: A credibility framework for effortless llm-powered fact checking. *IEEE Access*.

Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. 2023. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

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.

Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7:1341697.

Liam Van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. *arXiv preprint arXiv:2210.13210*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.