IRIT-Berger-Levrault at SemEval-2024: How Sensitive Sentence Embeddings are to Hallucinations?

Nihed Bendahman [◇] [♠], **Karen Pinel-Sauvagnat** [◇],
Gilles Hubert [◇], **Mokhtar Boumedyen Billami** [♠]
[◇]Université Paul Sabatier, IRIT, UMR 5505 CNRS, Toulouse, France
[♠]Berger-Levrault, 64 Rue Jean Rostand, 31670 Labège, France
{nihed.bendahman, karen.sauvagnat, gilles.hubert}@irit.fr
{nihed.bendahman, mb.billami}@berger-levrault.com

Abstract

This article presents our participation to Task 6 of SemEval-2024, named SHROOM (*a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes*), which aims at detecting hallucinations. We propose two types of approaches for the task: the first one is based on sentence embeddings and cosine similarity metric, and the second one uses LLMs (Large Language Model). We found that LLMs fail to improve the performance achieved by embedding generation models. The latter outperform the baseline provided by the organizers, and our best system achieves 78% accuracy.

1 Introduction

In recent years, with the emergence of the foundation models, the generation of hallucinated text has become an increasingly prominent and alarming issue. Despite the state-of-the-art performances achieved by the latest text generation models, such as GPT-4 (Achiam et al., 2023) or Llama 2 (Touvron et al., 2023), the problem of hallucinations remains open, making these models challenging to apply in real-world applications.

Hallucination is defined as a segment of text that appears fluent and natural but contains incoherent and inconsistent information compared to the provided context (Ji et al., 2023). The problem of hallucinations appears in several NLG tasks such as text summarization (Cao et al., 2021; Zhang et al., 2022) and machine translation (Xu et al., 2023). The shared-task Shroom[1] falls within the scope of these tasks.

The aim of Shroom is to identify samples containing hallucinations with regard to the provided context through a binary classification. The task is established in a post hoc setting, where models have already been trained to generate text based on the provided context. Three text generation tasks are considered: machine translation (MT), paraphrasing generation (PG), and definition modeling (DM). One can find in Table 1 a sample of data for the Machine Translation task. Participants should find the *label* of the *hypothesis* that was generated by the model, given the *target* or the *source*. For instance, here the aim is to determine if the hypothesis (*I've got the floor and the furniture.*) is a hallucination given the source (*J'ai poli le plancher et les meubles.*) or the target (*I polished up the floor and furniture.*). In this example, the hypothesis is labeled as hallucination (*label*) with regard to the assessments made by 3 annotators (*labels*).

Source :	J'ai poli le plancher et les meubles.
Target :	I polished up the floor and furniture.
Hypothesis :	I've got the floor and the furniture.
Ref :	Either
Labels :	[Hallucination, Hallucination, Hallucination]
Label :	Hallucination
p(hallucination) :	1.0

Table 1: Data sample

The binary classification can be performed in two different tracks: the model-aware and model-agnostic tracks. In the model-aware track, the checkpoint of the model that generated the hypothesis is provided and can be used in the classification system, which is not the case for the model-agnostic track. For more information on task 6 of SemEval-2024 Shroom as described by its organizers, we invite the reader to consult the paper by (Mickus et al., 2024).

In the literature, detecting hallucinations in sentences mainly relies on comparing segments of text. Among these approaches, one can cite those based on named entities (Nan et al., 2021): the idea is to determine if the generated text contains incon-

[1] https://helsinki-nlp.github.io/shroom/

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 573–578
June 20-21, 2024 ©2024 Association for Computational Linguistics

sistent entities compared to the source. Other approaches are based on question-answering methods, where the aim is to answer a set of questions and evaluate the difference between the two texts (Wang et al., 2020). However, data provided for the Shroom task are of a particular nature as it consists of very short texts, composed of one or two concise sentences. This characteristic limits the use of several detection methods, such as those based on named entities, since they are very rare in these texts. The same goes for question-answering methods. This observation has led us to turn to simpler methods, involving capturing the semantics of sentences in a general way.

In this paper, we present the two lines of approaches we investigated for the task:

- The first one is based on embedding models. We solve the hallucination classification task using the cosine similarity metric between sentence embeddings of the reference and hypothesis (see Section 4.1),
- The second one relies on Large Language models with specified prompts to classify the generated text that contains hallucinations. We use Llama 2 and Mistral using 2 different prompts inspired by SelfCheckGPT (Manakul et al., 2023).

Our sentence embedding approaches outperformed our LLM ones. The former have proven to be very relevant given the shortness of the sentences and the low risk of losing information. Our best performing system obtained the accuracy of 78% in both tracks which puts us in position 22/48 in the model-agnostic track and 22/45 in the model-aware track.

2 Related Work

Approaches in the literature can be classified depending on the hallucinatory content to detect (Ji et al., 2023). Some works focus on the detection and comparison of existing entities between the context and the generated hypothesis, such as (Nan et al., 2021). They are based on the assumption that human brain is sensitive to different types of information, such as named entities and proper nouns when reading, and mistakes concerning named entities are striking to human users (Ji et al., 2023). (Feng et al., 2023) go further by evaluating facts (entities and relations). Another line of works focuses on the use of question-answering as an indicator to identify hallucinations (Wang et al., 2020; Scialom et al., 2021), or the use of text entailment, which consists in determining whether the generated text is a hallucination if it cannot be entailed by the source (Falke et al., 2019).

Other approaches focused on the classification of hallucination types (whether they are intrinsic or extrinsic (Maynez et al., 2020)), or factual or non-factual (Cao et al., 2021).

Large Language Models have also been used to determine whether the generated text contains hallucinatory content. The aim of these methods is to set up prompts to compare the sources and the hypotheses (Manakul et al., 2023; Chern et al., 2023). Other methods make specific prompts to ask the LLM to "think" and judge whether a given text contains hallucinations, justifying its answer by producing a chain-of-thought (CoT) explanation (Friel and Sanyal, 2023).

In this paper, we explore a simpler method that consists in calculating the embeddings of the hypothesis and the reference, and computing their semantic similarity using the cosine similarity metric. Contextual embeddings have been successfully used in various NLP tasks, such as sentiment analysis (Carrasco and Dias, 2023) or topic modeling (Schneider, 2023). Our underlying idea here is to see how sensitive the semantic similarity is to the hallucinated content in sentences and to what extent the cosine similarity metric reflects this sensibility.

3 Data Description

The organizers of Shroom provided data from 3 different NLG tasks : Machine Translation, Paraphrasing Generation, Definition Modeling. Each task is divided into two tracks: model-aware track and model-agnostic track. We were provided 5 datasets in the development phase : train-aware, train-agnostic, trial, dev-aware and dev-agnostic, containing 30000, 30000, 80, 501 and 499 samples respectively and 2 different test datasets in the evaluation phase: test model-aware and test model-agnostic, containing each 1500 samples.

For each sample, the model-generated hypothesis was annotated by 3 (trial dataset) or 5 different annotators (dev and test datasets) (*labels* in Table 1). Annotators were asked to assess whether the generated hypothesis was consistent with the reference and to provide a label {hallucination, not-hallucination}. At the end of the annotation process, the most preponderant label is chosen as the final label (*label* in Table 1), with an assigned probability corresponding to the proportion of annotators who considered this specific datapoint to be an hallucination (*p(hallucination)* in Table 1).

4 System Overview

As the training dataset provided by organizers is not labeled, we decided to experiment unsupervised approaches, either using sentence embeddings or LLMs.

4.1 Embedding-Based Approach

We first generate the contextual embeddings of the reference and the hypothesis. For the paraphrasing generation task, we consider the source as the reference, while for the other two tasks, the target is taken as the reference. Next, we calculate the cosine similarity between these two embeddings. If this similarity does not exceed a predefined threshold, we assign the label "hallucination". We evaluated various embedding models namely Sentence-T5 XL (Ni et al., 2021), a specialized variant of the T5 model designed specifically to generate representations of sentences; BGE-base; BGE-large (Xiao et al., 2023); E5-base; E5-large and SF E5 (Wang et al., 2022). We compared their performances to determine their effectiveness in our task using different cosine similarity thresholds (see Section 5.2). This comparison enabled us to select the most appropriate model with the cosine threshold that maximizes the classification accuracy.

Participants were also asked to estimate a probability of the predicted labels. To estimate this probability, we apply an empirical rule. Let t be our threshold, $cossim$ the value of the cosine similarity, l our predicted label et $p(l)$ the probability of hallucination. Algorithm 1 details the rules we applied.

The idea behind this rule is that the further we are from the cosine similarity threshold, the more certain we are that the hypothesis generated by the language model is a hallucination.

4.2 LLM-Based Approach

We tested two LLMs, Llama-2-13b (Touvron et al., 2023) and Mistral-7b (Jiang et al., 2023), with 2 different prompts inspired by SelfCheckGPT (Manakul et al., 2023). This enabled us to make a direct comparison with the baseline system given by the organizers, which is based on a variant of Mistral fine-tuned with the same first prompt we used.

Algorithm 1 Hallucination Probability Estimation

```
if $cossim \geq t$ then
    $l \leftarrow Not\ Hallucination$
    if $cossim \leq t + \epsilon$ then
        $p(l) \leftarrow 0.33$
    else
        $p(l) \leftarrow 0.0$
    end if
else
    $l \leftarrow Hallucination$
    if $cossim \geq t - \epsilon$ then
        $p(l) \leftarrow 0.66$
    else
        $p(l) \leftarrow 1.0$
    end if
end if
```

As our method uses only the generated hypotheses and the references to detect hallucinations, we used a single model for both model-aware and model-agnostic tracks.

5 Experiments and Results

5.1 Experimental Setup

For all the models used, we retrieved the checkpoints from the HuggingFace website [2].

- For the embedding-based approach, we experimentally fix ϵ to 0.05 (see Algorithm 1) using the dev-set. t is also fixed experimentally. Experiments conducted to determine its value are detailed in section 5.2.

- For the LLMs approach, we use the Langchain framework [3] to set up the prompts and query the LLMs. Table 2 describes the two prompts we used. Prompt 1 is directly taken from SelfCheckGPT's system (Manakul et al., 2023) which serves as Baseline. The idea of the work of (Manakul et al., 2023) is to ask the LLM an explicit and simple question. They show that with this kind of prompts, LLMs better understand the task they are asked to perform. With regard to prompt 2, we wanted to experiment whether introducing the concept of "hallucination" in the prompt and specifying its definition helps the LLM better classify.

[2] hhttps://huggingface.co/

[3] https://www.langchain.com/

Prompt 1 (Manakul et al., 2023)	Context:{} Sentence:{} Is the Sentence supported by the Context above? Answer using ONLY yes or no:
Prompt 2	Context:{} Sentence:{} Is the Sentence a hallucination (which means it contains inconsistent or incoherent information) compared to the Context above? Answer using ONLY yes or no:

Table 2: Prompts used to perform the binary classification.

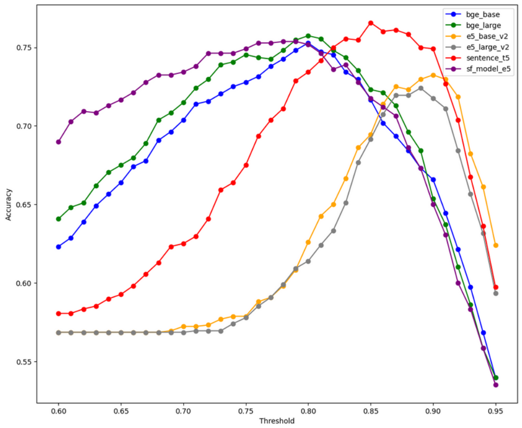

Figure 1: Variation of the threshold of the cosine similarity metric maximizing the accuracy of the classification system as a function of the models over the Trial and Dev datasets.

5.2 Preliminary Experiments

In this section, we describe the experiments made to determine the threshold of the cosine similarity metric that maximizes classification accuracy (t in Algorithm 1). We defined an interval ranging from 0.6 to 0.95. Then, for each value of the interval by step of 0.01, we performed the classification and calculated its accuracy. This experimentation was conducted on the 3 labelled datasets, namely trial, dev-aware and dev-agnostic. The graph in Figure 1 shows the evolution of accuracy as a function of the cosine similarity threshold used to define hallucination. We can see that the models behave in a similar way: accuracy rises progressively with the threshold values used, reaching a peak around the [0.78, 0.9] interval. We can also see that for the models for which we used two variants as BGE-base, BGE-large as well as E5-base and E5-large, the behavior of the variants is almost identical with a few small differences. Table 3 reports the selected threshold for each model applied on the test dataset in the evaluation phase.

5.3 Results

Two evaluation metrics are used for the task: the accuracy of the classification and the Spearman correlation of the system's output probabilities with the proportion of the annotators marking the item as hallucinated. The official ranking was made on the basis of the accuracy, with a tie-breaker between systems having obtained the same score using the Spearman correlation. As we did not provide classification probabilities for the LLMs approach, we only report them for the embedding-based approach.

Table 3 shows the results we obtained with the different embedding models submitted. We can see that all the models exceed the baseline on the two metrics used, with the best performance coming from the Sentence-T5 model. Given that the baseline consists in the use of an LLM with a prompt, we can say that the embedding models used with the right threshold distinguish fairly well between hallucinated and non-hallucinated hypotheses compared to LLM with the used prompts. Since organizers published official results and released the test sets, we re-ran our experiments varying t, threshold used with the cosine metric. The results, not reported here, are consistent with those of the trial and dev collections. This leads us to believe that our approach of threshold selection is robust.

Table 4 shows the results obtained with the LLMs we used. We can see that they do not perform as well as the embedding models, and do not exceed the baseline. Regarding the prompts we used, no conclusion can be drawn for the moment. Further experiments are required.

With the scores obtained by the sentence-T5 model, we were ranked 22/45 and 22/48 in the model-aware and model-agnostic tracks respectively. It is worth noting that the first half of the ranking is extremely tight. It often takes 4 decimal digits to separate the accuracy of the various participants.

6 Conclusion

In this paper, we summarize our participation to task 6 of the SemEval-2024 evaluation campaign:

Model	Value of t for Aw	M-Aw	SC-Aw	Value of t for Ag	M-Ag	SC-Ag
Baseline	/	0.745	0.487	/	0.696	0.402
Best system	/	0.812	0.699	/	0.847	0.769
Worst system	/	0.483	-0.06	/	0.460	0.133
BGE-Base	0.77	0.750	0.552	0.79	0.754	0.563
BGE-Large	0.77	0.766	0.569	0.78	0.766	0.581
E5-Base	0.87	0.742	0.494	0.90	0.748	0.531
E5-Large	0.87	0.754	0.510	0.89	0.751	0.525
Sentence T5 XL	0.86	**0.781**	**0.601**	0.85	**0.782**	**0.636**
SF E5	0.75	0.758	0.523	0.79	0.762	0.540

Table 3: Results obtained (accuracy (M) and Spearman correlation (SC)) with each embedding model using the selected threshold, in comparison to the Baseline, best and worst submitted systems. Results are reported for the model-aware (Aw) and model-agnostic (Ag) tracks.

Model	M-Aw	M-Ag
Llama-2-13b-chat Prompt 1	0.618	0.557
Llama-2-13b-chat Prompt 2	0.555	0.536
Mistral-7b-instruct Prompt 1	0.627	0.519
Mistral-7b-instruct Prompt 2	0.676	0.618

Table 4: Results obtained (accuracy M) for each LLM with the two prompts used. Results are reported for the model-aware (Aw) and model-agnostic (Ag) tracks.

Shroom. We presented two different approaches: one based on the use of embedding models with a cosine similarity threshold to perform the binary classification, and the other based on LLM using simple prompts to detect hallucinatory content. We showed that on the data used for the task, embedding generation models perform better than LLMs. In future work, we will explore this approach a little further, by fine-tuning the models used for instance.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014740 made by GENCI.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2021. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *arXiv preprint arXiv:2109.09784*.

Paulo Carrasco and Sandra Dias. 2023. Exploring natural language processing and sentence embeddings for sentiment analysis of online restaurant reviews. *Atas da 23ª Conferência da Associação Portuguesa de Sistemas de Informação*.

Ethan Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai - a tool augmented framework for multi-task and multi-domain scenarios. *ArXiv*, abs/2307.13528.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Annual Meeting of the Association for Computational Linguistics*.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. *ArXiv*, abs/2305.08281.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *ArXiv*, abs/2310.18344.

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.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv*, abs/2303.08896.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. *ArXiv*, abs/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)*, Mexico City, Mexico. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cícero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Matthew Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *ArXiv*, abs/2108.08877.

Johannes Schneider. 2023. Efficient and flexible topic modeling using pretrained embeddings and bag of sentences. *ArXiv*, abs/2302.03106.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. In *Conference on Empirical Methods in Natural Language Processing*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *ArXiv*, abs/2004.04228.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *ArXiv*, abs/2309.07597.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, Kazuma Hashimoto, and Yingbo Zhou. 2022. Improving the faithfulness of abstractive summarization via entity coverage control. *arXiv preprint arXiv:2207.02263*.