BrainLlama at SemEval-2024 Task 6: Prompting Llama to detect hallucinations and related observable overgeneration mistakes

Marco Siino
Department of Electrical, Electronic
and Computer Engineering
University of Catania
Italy
marco.siino@unipa.it

Abstract

Participants in the SemEval-2024 Task 6 were tasked with executing binary classification aimed at discerning instances of fluent overgeneration hallucinations across two distinct setups: the model-aware and model-agnostic tracks. That is, participants must detect grammatically sound output which contains incorrect or unsupported semantic information, regardless of whether they had access to the model responsible for producing the output or not, within the model-aware and model-agnostic tracks. Two tracks were proposed for the task: a model-aware track, where organizers provided a checkpoint to a model publicly available on HuggingFace for every data point considered, and a model-agnostic track, where the organizers do not. In this paper, we discuss the application of a Llama model to address both the tracks. Our approach reaches an accuracy of 0.62 on the agnostic track and of 0.67 on the aware track.

1 Introduction

In the modern Natural Language Generated (NLG) domain, two interconnected challenges persist: neural models often produce linguistically fluent, yet inaccurate, output, while evaluation metrics primarily focus on fluency rather than accuracy. This situation leads to the phenomenon of "hallucinations," wherein neural networks generate output that sound plausible but deviate from the intended meaning, posing difficulties in automatic detection. However, in many NLG applications, the accuracy of output is paramount. For instance, generating translations that diverge from the source text undermines the effectiveness of machine translation systems. Also, as reported in recent survey papers, LLMs are prone to hallucinations, as proven in a variety of recent survey papers (Huang et al., 2023; Ji et al., 2023; Zhang et al., 2023). This LLMs drawback led to the proposal of SemEval-2024 Task 6 (Mickus et al., 2024), where participants were tasked with conducting detection of hallucinations across two subtracks: model-agnostic and model-aware. Put simply, participants were required to detect grammatically correct output containing incorrect or unsupported semantic information, regardless of access to the model responsible for generating them. In the literature, the task has been recently addressed with prompt engineering strategies that provide further context to the models to properly drive and control the models' output (Martino et al., 2023; Li et al., 2024).

To aid in this assignment, a dataset including references, inputs, checkpoints, and outputs from systems trained for three NLG tasks (definition, modeling, machine translation, and paraphrase generation) was provided. These systems were trained with varying levels of accuracy. The dataset included development and test sets annotated by a minimum of five annotators, with a majority vote establishing the gold label for binary annotations.

To address these objectives, there is an ongoing demand for automated tools capable of extracting and categorizing data, facilitating the classification of NLG content containing hallucinations. Recent advancements in machine and deep learning architectures have spurred heightened interest in Natural Language Processing (NLP). Substantial endeavors have been directed towards devising techniques for the automated identification and categorization of textual content accessible on the internet today. In the literature, to perform text classification tasks, several strategies have already been proposed (Kim, 2014; Siino et al., 2024a; Lomonaco et al., 2023).

To face with the task, we propose a Transformer-based approach which made use of Llama (Touvron et al., 2023). We used the model in a zero-shot setup described in the rest of this paper. Specifically, we prompted the latest pre-trained version of Llama with any sample in the dataset. Specifically, we provided a *context* and a *sentence*, asking the

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 82–87
June 20-21, 2024 ©2024 Association for Computational Linguistics

model if the sentence was really supported by the context or was an example of hallucination.

The subsequent sections of the paper are structured as follows: Section 2 offers background information on Task 6, held at SemEval-2024. In Section 3, we outline the approach introduced in this study. Section 4 delves into the specifics of the experimental setup employed to reproduce our findings. The outcomes of the official task and relevant discussions are presented in Section 5. Finally, Section 6 concludes our study and suggests avenues for future research.

We make all the code publicly available and reusable on GitHub[1].

2 Background

This section furnishes background information regarding Task 6, held at SemEval-2024 (named, *SHROOM*). SHROOM participants are tasked with identifying grammatically correct output containing incorrect semantic information, regardless of their access to the model responsible for generating the output.

The data files are formatted as JSON lists, with each element representing a datapoint. Each datapoint corresponds to a different model production and includes the following details:

- Task (task): indicating the objective the model was optimized for.
- Source (src): the input provided to the models for the generation.
- Target (tgt): the intended reference "gold" text that the model should generate.
- Hypothesis (hyp): the actual output generated by the model.
- Annotator labels (labels): indicating whether each individual annotator considered this datapoint to be a hallucination or not.
- Majority-based gold label (label): based on the previous per-annotator labels.
- Probability of hallucination (p(Hallucination)): representing the proportion of annotators who deemed this specific datapoint to be a hallucination.
- Indicator of semantic reference (ref): specifying whether the target, source, or both contain the semantic information necessary to determine if a datapoint is a hallucination.

Furthermore, model-aware datapoints also identify the model used to produce each datapoint, represented by a Hugging Face identifier (model).

For each sample in the dataset, there is a source text, a target text and a hypothesis text. Depending on the task (DM, MT, PG) the goal is to determine if the Hypothesis contains any hallucination.

In the Table 1 there are three different samples from the official test set. Even if the labels are shown in the table along with the hallucination probabilities, during the evaluation phase of the competition, labels, and probabilities were hidden for the participants.

3 System Overview

Even if it has already been proved that the Transformers are not necessarily the best option for every text classification task (Siino et al., 2022), depending on the goal some strategies like domain-specific fine-tuning (Sun et al., 2019; Van Thin et al., 2023), or data augmentation (Lomonaco et al., 2023; Mangione et al., 2022; Siino et al., 2024a) can be beneficial for the considered task.

However, to address the task 6 hosted at SemEval-2024, we made use of a zero-shot learning approach (Chen et al., 2023; Wahidur et al., 2024), making use of the GPT Transformer named Llama 7B. This was dictated by our choice to bear in mind the computational efficiency without further feature engineering and/or heavy data preprocessing strategies.

Llama 2, a suite of large language models (LLMs), includes pretrained and fine-tuned models ranging from 7 to 70 billion parameters. Specifically tailored for dialogue applications, the fine-tuned LLMs are designated as Llama 2-Chat. The models demonstrate interesting performance when compared to open-source chat models across the majority of assessed benchmarks. Additionally, according to human evaluations focusing on helpfulness and safety, they could potentially serve as viable substitutes for closed-source models. Even if several others Open LLMs have proved to be able of outperforming Llama (Jiang et al., 2023), here we investigate the model's actual performance on this specific task. The authors of the model of-

[1] https://github.com/marco-siino/SemEval2024/

Target Text	Hypothesis Text	Label	p(Hallucination)
"Would you be surprised if I told you my name isn't actually Tom?"	*"You're gonna be surprised if I say my real name isn't Tom?"*	Not Hallucination	0.0
"There will be plenty of food."	*"The food will be full."*	Hallucination	0.8
"The two brothers are pretty different."	*"There's a lot of friends."*	Hallucination	1.0

Table 1: Three samples from the official test set are provided. Together with the labels for each sample, is also reported the probability of hallucination.

fer a comprehensive account of the fine-tuning approach and safety enhancements for Llama 2-Chat, with the aim of facilitating community engagement and contributing to the responsible advancement of LLM technology.

The Llama 2 suite comprises:

- Llama 2: an enhanced iteration of Llama 1, trained on a revised assortment of publicly available data. Notable improvements include a 40% augmentation in the size of the pretraining corpus, a doubling of the model's context length, and the adoption of grouped-query attention. Variants of Llama 2 with 7 billion, 13 billion, and 70 billion parameters are being released. Additionally, authors have trained 34 billion parameter variants, detailed in their paper but not released to the public;
- Llama 2-Chat: a fine-tuned version of Llama 2 tailored for dialogue applications.

To develop the new Llama 2 model family, the authors commenced with the pretraining methodology outlined in Touvron et al. 2023, utilizing an optimized autoregressive transformer. However, the authors made several modifications to enhance performance. These included more rigorous data cleaning, updates to data mixtures, training on 40% more total tokens, doubling the context length, and implementing grouped-query attention (GQA) to enhance inference scalability, particularly for larger models.

More specifically, given the task hosted at SemEval-2024, we asked the model: *"Is the Sentence supported by the Context above? Answer using ONLY yes or no:"*. To this request, the model replied with one or more words — usually starting with *yes* or *no* — that we parsed to extract one of the two labels. For example, given the context:

> *"The East African Islands are in the Indian Ocean off the eastern coast of Africa"*

The sentence:

> *"The eastern islands of the Indian Ocean are located in the eastern part of the Indian Ocean"*

And our question:

> *Is the Sentence supported by the Context above? Answer using ONLY yes or no:*

The model replied with:

> *no, the sentence is not supported by the context provided*

that we mapped into the label *Hallucination*.

It is worth noting that we needed to post-process the model answers to extract only the first word of the reply (i.e., *yes* or *no*). The model barely replied with a single word, even if prompted with the specific request of limiting its answer.

In the literature, several prompt engineering strategies have already been introduced (Denny et al., 2023; Giray, 2023). However, also from this perspective, we opted for a straight interaction with the GPT model, without any further engineering of the process. Finally, we collected all the predictions provided on the test set to into a JSON file with the required format to submit our predictions.

As noted in the recent study by Siino et al. 2024b, the contribution of preprocessing for text classification tasks is generally not impactful when using Transformers. More specifically, the best combination of preprocessing strategies is not very different from doing no preprocessing at all in the case of Transformers. For these reasons, and to keep our system highly fast and computationally light, we have not performed any preprocessing on the text.

4 Experimental Setup

We implemented our model on Google Colab. The library we used come from HuggingFace and as already mentioned is Llama 2[2]. Llama 2 comprises a series of pretrained and fine-tuned generative text models with parameter ranges spanning from 7 billion to 70 billion. This repository specifically hosts the 7B fine-tuned model, tailored for dialogue applications and converted to the Hugging Face Transformers[3] format. We also imported the Llama library (Touvron et al., 2023) from *llama_cpp*. The library is fully described on GitHub[4]. The dataset provided for all the phases are available on the Official Competition page. We did not perform any additional fine-tuning on the model. To run the experiment, a T4 GPU from Google has been used. After the generation of predictions, we exported the results on the format required by the organizers. As already mentioned, all of our code is available on GitHub.

5 Results

Submissions were divided into two tracks: a model-aware track, where organizers provide a checkpoint to a model publically available on Hugging Face for every data point considered, and a model-agnostic track, where organizers do not. The organizers encouraged participants to make use of model checkpoints in creative ways. For both tracks, all participants' submissions were evaluated using two criteria: the accuracy that the system reached on the binary classification; and the Spearman correlation of the systems' output probabilities with the proportion of the annotators marking the item as overgenerating. The evaluation script was made available[5], along with baseline systems and format checkers.

In the Table 2 we report the results obtained by our approach. In the rows are reported the two tracks (i.e., model agnostic or model aware) while in the column are reported the results according to the output score provided on CodaLab. As can be noted from the Tables 3, 4 our proposed approach it is not able to outperform the baseline provided for the task (i.e., Mistral 7B).

In the Table 3 and in the Table 4, the results obtained by the first three teams and by the last one, as showed on the official task page, are reported. Compared to the best performing models, our simple approach exhibits some room for improvements. Furthermore, our proposed approach is not able to outperform the baseline provided for the task. For this reason, we are confident that no further investigations should be performed for this task making use of the Llama model. However, it is worth notice that it required no further pre-training and the computational cost to address the task is manageable with the free online resources offered by Google Colab.

	Acc	Rho
Agnostic	0.625	0.204
Aware	0.671	0.244

Table 2: The method's performance on the test set. In the table are reported the results obtained by our private area on CodaLab.

6 Conclusion

This paper presents the application of a Llama-model for addressing the Task 6 at SemEval-2024. For our submission, we decided to follow a zero-shot learning approach, employing as-is, an in-domain pre-trained Transformer. After several experiments, we found beneficial to build a prompt containing the question for the model. Then we provide as a prompt the target sentence and the hypothesis sentence. The model was asked to decide if the hypothesis sentence is supported by the content of the target sentence, or if it is just a hallucinated text. The task is challenging, and there is still opportunity for improvement, as can be noted looking at the final ranking. Possible alternative approaches include utilizing the few-shot capabilities or also the use of other models like GPT and T5, increasing the size of the training set by using further data, or directly integrating other samples from the training and from the development sets. Further improvements could be obtained with a fine-tuning and modelling the problem as a text classification task. Furthermore, given the interesting results recently provided on a plethora of tasks, also other few-shot learning (Wang et al., 2023; Maia et al., 2024; Siino et al., 2023; Meng et al., 2024) or data augmentation strategies (Muftie and Haris, 2023; Tapia-Téllez and Escalante, 2020; Siino and Tinnirello, 2023) could be employed to improve the

[2] https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGUF

[3] https://huggingface.co/

[4] https://github.com/ggerganov/llama.cpp

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

TEAM NAME	ACC	RHO
GroupCheckGPT (1)	0.847	0.769
OPDAI (2)	0.836	0.732
HIT_WL (3)	0.831	0.768
baseline system	0.697	0.403
OxYuan (48)	0.461	0.134

Table 3: Comparing performance on the test set for the model agnostic track. In the table are shown the results obtained by the first three teams and by the last one. In parentheses is reported the position in the official final ranking.

TEAM NAME	ACC	RHO
HaRMoNEE (1)	0.813	0.699
GroupCheckGPT (2)	0.806	0.715
TU Wien (3)	0.806	0.707
baseline system	0.745	0.488
octavianB (45)	0.483	-0.064

Table 4: Comparing performance on the test set for the model aware track. In the table are shown the results obtained by the first three users and by the last one. In parentheses is reported the position in the official final ranking.

results. Looking at the final ranking, our simple approach exhibits some room for improvements. However, it is worth notice that required no further pre-training and the computational cost to address the task is manageable with the free online resources offered by Google Colab.

Acknowledgments

We extend our gratitude to the anonymous reviewers for their insightful comments and valuable suggestions, which have significantly enhanced the clarity and presentation of this paper.

References

Shiming Chen, Ziming Hong, Wenjin Hou, Guo-Sen Xie, Yibing Song, Jian Zhao, Xinge You, Shuicheng Yan, and Ling Shao. 2023. Transzero++: Cross attribute-guided transformer for zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12844 – 12861.

Paul Denny, Viraj Kumar, and Nasser Giacaman. 2023. Conversing with copilot: Exploring prompt engineering for solving cs1 problems using natural language. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 1136–1142.

Louie Giray. 2023. Prompt engineering with chatgpt: a guide for academic writers. *Annals of biomedical engineering*, 51(12):2629–2633.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

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 Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

Francesco Lomonaco, Marco Siino, and Maurizio Tesconi. 2023. Text enrichment with japanese language to profile cryptocurrency influencers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2708–2716. CEUR-WS.org.

Beatriz Matias Santana Maia, Maria Clara Falcão Ribeiro de Assis, Leandro Muniz de Lima, Matheus Becali Rocha, Humberto Giuri Calente, Maria Luiza Armini Correa, Danielle Resende Camisasca, and Renato Antonio Krohling. 2024. Transformers, convolutional neural networks, and few-shot learning for classification of histopathological images of oral cancer. *Expert Systems with Applications*, 241:122418.

Stefano Mangione, Marco Siino, and Giovanni Garbo. 2022. Improving irony and stereotype spreaders detection using data augmentation and convolutional neural network. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2585–2593. CEUR-WS.org.

Ariana Martino, Michael Iannelli, and Coleen Truong. 2023. Knowledge injection to counter large language model (llm) hallucination. In *European Semantic Web Conference*, pages 182–185. Springer.

Zong Meng, Zhaohui Zhang, Yang Guan, Jimeng Li, Lixiao Cao, Meng Zhu, Jingjing Fan, and Fengjie Fan. 2024. A hierarchical transformer-based adaptive metric and joint-learning network for few-shot rolling bearing fault diagnosis. *Measurement Science and Technology*, 35(3).

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.

Fuad Muftie and Muhammad Haris. 2023. Indobert based data augmentation for indonesian text classification. In *2023 International Conference on Information Technology Research and Innovation, ICITRI 2023*, page 128 – 132.

Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, and Marco La Cascia. 2022. Fake news spreaders detection: Sometimes attention is not all you need. *Information*, 13(9):426.

Marco Siino, Francesco Lomonaco, and Paolo Rosso. 2024a. Backtranslate what you are saying and i will tell who you are. *Expert Systems*, n/a(n/a):e13568.

Marco Siino, Maurizio Tesconi, and Ilenia Tinnirello. 2023. Profiling cryptocurrency influencers with few-shot learning using data augmentation and ELECTRA. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2772–2781. CEUR-WS.org.

Marco Siino and Ilenia Tinnirello. 2023. Xlnet with data augmentation to profile cryptocurrency influencers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2763–2771. CEUR-WS.org.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024b. Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, 121:102342.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer.

José Medardo Tapia-Téllez and Hugo Jair Escalante. 2020. Data augmentation with transformers for text classification. In *Advances in Computational Intelligence*, pages 247–259, Cham. Springer International Publishing.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Dang Van Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023. Vietnamese sentiment analysis: An overview and comparative study of fine-tuning pretrained language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–27.

Rahman S. M. Wahidur, Ishmam Tashdeed, Manjit Kaur, and Heung-No Lee. 2024. Enhancing zero-shot crypto sentiment with fine-tuned language model and prompt engineering. *IEEE Access*, 12:10146 – 10159.

Xixi Wang, Xiao Wang, Bo Jiang, and Bin Luo. 2023. Few-shot learning meets transformer: Unified query-support transformers for few-shot classification. *IEEE Trans. Circuits Syst. Video Technol.*, 33(12):7789–7802.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.