Two Stage Psychology-Guided Fine-Grained Editing and Sampling Approach for Mitigating Hallucination in Large Language Models

Lei Chen (202322090726@std.uestc.edu.cn)
Xiaohua Wu* (wuxh@uestc.edu.cn)
Zihan Xiong (2022090905007@std.uestc.edu.cn)
Xuanshuo Kang (202422090729@std.uestc.edu.cn)
School of Information and Software Engineering, University of Electronic Science and Technology of China
Chengdu, China

Abstract

The hallucination issue in large language models (LLMs) significantly restricts their application in high-stakes domains such as healthcare, cognitive science and law. Existing approaches primarily focus on data optimization or decoding strategies but lack a fine-grained analysis of the underlying mechanisms of hallucinations. This paper proposes a psychology-guided two-stage fine-grained editing and sampling framework (PGFES), which, for the first time, introduces psychological classifications of hallucinations into LLM optimization. Firstly, an attention-augmented MLP probe is designed to identify "truthfulness directions" corresponding to different hallucination types through feature channel reweighting, enabling fine-grained editing of the model's internal representations during inference. Then, a dynamic weighting mechanism based on Jaccard similarity is employed to compute the weights of multi-path edited outputs, achieving adaptive sampling. Experiments demonstrate that the optimization method incorporating psychology-related concepts improves truthfulness by 20.4% on the TruthfulQA open-domain question-answering task compared to baseline models and exhibits strong generalization across cross-domain datasets.

Keywords: LLMs; psychology-guided; attention-augmented MLP; fine-grained editing; sampling

Introductions

Current large language models (LLMs), such as GPT-4(Achiam et al., 2023), LLaMA(Touvron et al., 2023), and Mistral(Jiang et al., 2023), have been widely applied across various domains. Leveraging their powerful text generation capabilities, these models excel in natural language processing tasks, enabling the automated generation of high-quality text and bringing significant convenience to production and daily life. From content creation and code generation to customer service(M. Chen et al., 2021);(Thoppilan et al., 2022), the extensive application of LLMs has greatly improved efficiency and reduced labor costs. However, despite their numerous benefits, these models also exhibit notable limitations(Yao et al., 2024);(Tonmoy et al., 2024).

As tools created by humans, LLMs inevitably inherit certain human flaws, with one of the most concerning issues being the phenomenon of "hallucination"(Banerjee, Agarwal, & Singla, 2024). Traditionally defined, hallucination refers to instances where the model generates incorrect, overly subjective, causing significant confusion for users. In high-reliability domains,such as finance(Zhou et al., 2024), medicine(Ahmad, Yaramis, & Roy, 2023), and law(Dahl, Magesh, Suzgun, & Ho, 2024), decision-making accuracy is critical, and erroneous advice can not only disrupt business operations but also lead to substantial economic losses or even endanger lives(Zhang et al., 2023);(Huang et al., 2025).

To alleviate hallucination in LLMs, scholars have proposed various methods, including the construction of high-quality datasets(Rejeleene, Xu, & Talburt, 2024), unbiased fine-tuning, editing hidden layers during inference(Orgad et al., 2024), incorporating external knowledge bases(J. Li, Yuan, & Zhang, 2024), and chain-of-thought prompting(Zheng et al., 2024). These methods have improved the authenticity of model outputs to some extent, but most have overlooked the perspective of addressing hallucination from its conceptual foundation. The two key points are as follows:

- Despite existing fine-grained categorizations of LLM hallucinations, there is a notable lack of research leveraging psychological concepts for such classification to specifically enhance response authenticity.
- Previous research employed multi-round output sampling and Chain-of-Thought to enhance model response authenticity. However, these methods' sampling relies on preset temperature/prompts and predominantly focuses only on response consistency.

Methods to enhance the authenticity of LLM responses can be divided into three stages: training, inference, and post-inference(Cao, Lin, Han, & Sun, 2024). The training stage primarily relies on task-specific datasets for downstream tasks, making this approach suitable for optimizing models for specific tasks. This paper focuses on the inference and post-inference stages to effectively mitigate hallucination without relying on large amounts of new data.

This paper proposes a hallucination mitigation framework, termed the Two Stage Psychology-Guided Fine-Grained Editing and Sampling Method (PGFES), which reduces hallucination through knowledge editing and targeted output sampling. Specifically, we utilize trained probes to identify fact-sensitive layers and attention heads that guide truthful information output. Based on this, we calculate fine-grained "truthful" directions corresponding to hallucination. During the inference stage, we edit internal representations along different fine-grained "truthful" directions to generate a batch of edited outputs. In the second stage, we compute the consistency weights of these edited outputs and select the output

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

with the highest consistency as one of the sampling references to produce the final response. Experimental results show that this method achieved a 20.4% improvement over the baseline score on the TruthfulQA dataset. The main contributions of this paper are as follows:

- We are the first to apply the psychological concept of hallucination to knowledge editing in LLMs. By integrating attention-based probes to obtain fine-grained "truthful" directions for different types of hallucination, we enhance existing editing methods.
- Based on fine-grained hallucination-edited responses, we propose a dynamic mechanism for calculating sampling reference weights, leveraging multi-perspective fine-grained editing to generate the final output. This method significantly improves the interpretability of sampled responses and enhances self-consistency.
- Our experiments demonstrate that the adoption of a two-stage psychology-guided fine-grained hallucination editing approach, combined with a sampling output strategy, significantly enhances the factual accuracy of LLMS on open-ended generation datasets, proving the validity of using psychology as a guide for hallucination classification in LLMs.

Related Works

In previous research, approaches to enhance the authenticity of large language models (LLMs) can be broadly categorized into two directions based on the stage of intervention: Reducing hallucinations during inference and Reducing hallucinations post-inference.

Methods to reduce hallucinations during the inference stage primarily include knowledge editing, decoding strategy optimization. These methods play a crucial role in ensuring the factual accuracy of the content generated by LLMs. For instance, Factuality Enhanced Method(Lee et al., 2022) proposed a fact-core-based sampling algorithm that dynamically adjusts externally introduced fact cores during the generation process to guide the output. Previous studies (Burns, Ye, Klein, & Steinhardt, 2023); (Maiorca et al., 2023) have experimentally discovered that for the same question, feeding factual and erroneous inputs to LLMs for activation and feature extraction enables the models to learn latent knowledge. This finding indicates that the intermediate-layer activations of LLMs are correlated with factual outputs. Building on this, scholars introduced Inference-Time Intervention (ITI)(Lee et al., 2022), which edits for truthfulness within the attention heads of LLMs. However, this method employs a linear binary classifier as a probe, resulting in limited accuracy and modest improvements in truthfulness. Other researchers found that in the early layers of large language models, more attention is given to low-level information such as morphemes, while the later layers focus more on semantic truthfulness. Based on this observation, they proposed the Decoding by Contrasting Layers method(DoLa)(Chuang et al., 2024), which enhances the focus on factual information by contrasting the logits of early and final layers.

Post-inference methods to enhance factual accuracy include multi-round sampling, prompt-based refinement, and chain-of-thought approaches. Dr.Chen first proposed a multi-sample method in their USC(X. Chen et al., 2023) paper, where multiple sampling results and prompt templates are concatenated to generate the final response. This method primarily focuses on the self-consistency of multiple responses. Subsequently, several scholars pioneered the application of the chain-of-thought method to mitigate hallucinations in large models. This approach relies on self-refine(Madaan et al., 2023) to regenerate answers, but its effectiveness depends on the quality of each step in the chain-of-thought. The ID method (Cheng et al., 2024) is an improvement upon the USC method. It achieves this by setting varying sampling temperatures to predict the next token, and then making judgments and selections at each token generation step to produce the final response. Although this method enhances the efficiency of generation, it still requires a relatively high number of inference generations.

In this paper, we build upon the foundation of fine-grained hallucination classification guided by psychology. By introducing fine-grained psychology-informed hallucination editing during the reasoning phase and integrating multi-perspective results to guide the LLM's reflection in the post-reasoning phase, we aim to mitigate the generation of hallucinations by the LLM.

Methods

To mitigate hallucination phenomena in large language models, this paper proposes an optimization method that combines psychology-guided fine-grained hallucination editing with adaptive sampling. Figure 1 illustrates the overall workflow of our proposed PGFES approach.

Psychology-guided fine-grained hallucination editing directions

To obtain a psychology-guided fine-grained hallucination editing direction, we first perform manual annotation and classification of responses in the dataset. This process involves a psychological analysis of hallucinations in LLMs. Previous research(Berberette, Hutchins, & Sadovnik, 2024) has proposed using psychological terminology to describe hallucinations in LLMs, categorizing them into the following types: Source Amnesia, Recency Effect, Availability Heuristics, Suggestibility, Cognitive Dissonance, and Confabulation. The first three fine-grained hallucination types defined in this paper are exclusively related to the pre-trained data of LLMs and cannot be altered through inferencing editing. Therefore, we categorize the hallucinations in the dataset into three types: Confabulation, Suggestibility, and Cognitive Dissonance. Confabulation refers to the inclusion of incorrect information in the LLMs' responses. Suggestibility indicates that the LLMs are overly faithful to the user's

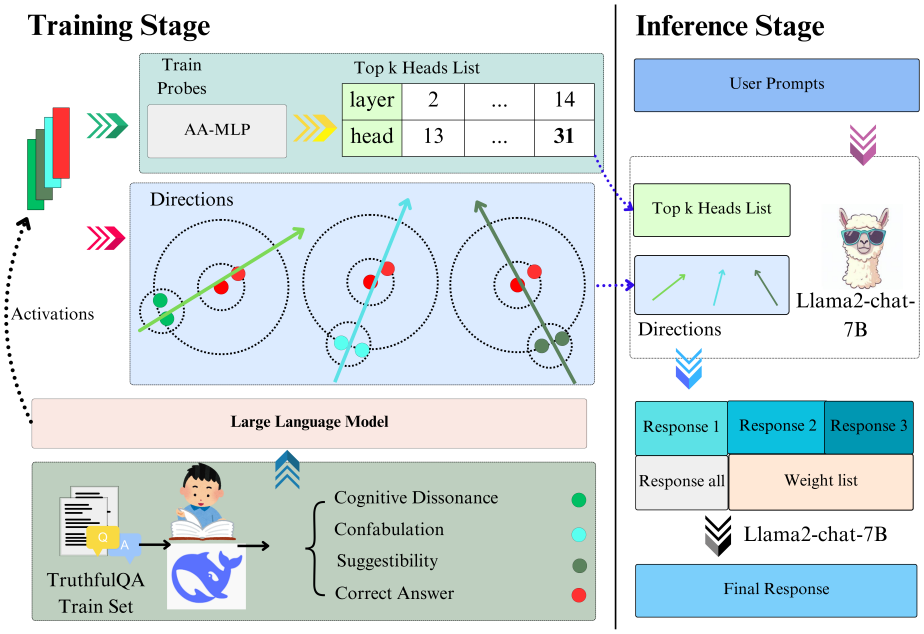

Figure 1: PGFES two-stage framework

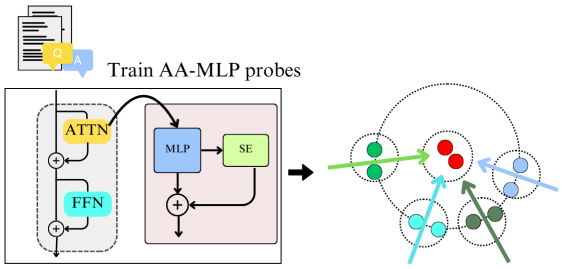

Figure 2: Calculate fine-grained directions

prompt input. Cognitive Dissonance denotes responses generated by the LLMs that conflict with the context. The annotation of the question-and-answer data is conducted in two steps. First, each incorrect response is manually labeled with its corresponding fine-grained hallucination type by human annotators. Subsequently, the manually annotated results from the same domain are used as few-shot examples to complete the annotation of the remaining data using the DeepSeek API. The annotation quality is ensured through human review. The datasets are defined as :$D_1 = \{Q, A_{cor}, H_{conf}\}$, $D_2 = \{Q, A_{cor}, H_{cog}\}$, $D_3 = \{Q, A_{cor}, H_{sug}\}$ and $D_4 = \{Q, A_{cor}, H_{all}\}$, where H_{all} represents all hallucinated responses without fine-grained classification, which are used to extract activation values and train the Attention-Augmented MLP. Unlike standard MLPs, we incorporate the SE Block (Squeeze-and-Excitation Block) into the MLP to compute channel-wise weights and deeply extract features.

The training of AA-MLP probes is formulated as a classification task. Specifically, for different psychological hallucination types, the outputs of each attention head in the LLM are extracted as inputs to the MLP, with the correctness of the corresponding prompt serving as the training labels. Ultimately, we obtain three single-hallucination probes and one all-hallucination probe. In the Squeeze-and-Excitation module(Hu, Shen, & Sun, 2018), we first perform the Squeeze operation by compressing input features through global average pooling, as defined by:

$$s_c = \frac{1}{d_{\text{out}}} \sum_{i=1}^{d_{\text{out}}} z_i^{(c)} \tag{1}$$

Next, the Excitation operation is performed, where features are first compressed through a dimensionality-reduction fully connected layer and then restored to their original dimensions through a dimensionality-increasing fully connected layer:

$$u = \sigma(W_{\text{down}} s + b_{\text{down}}), \quad W_{\text{down}} \in \mathbb{R}^n, \quad b_{\text{down}} \in \mathbb{R}^r. \tag{2}$$

$$a = \sigma(W_{\text{up}} u + b_{\text{up}}), \quad W_{\text{up}} \in \mathbb{R}^{d_{\text{out}} \times \frac{d_{\text{out}}}{r}}, \quad b_{\text{up}} \in \mathbb{R}^{d_{\text{out}}} \tag{3}$$

Finally, the attention weights are multiplied with the original feature channels to obtain weighted features: $\bar{x}_c = a_c \cdot x_c$

$$p_\theta(x_l^h) = \text{AA-MLP}\left(\left\langle \theta, x_l^h \right\rangle\right). \tag{4}$$

Here, θ is a set of trainable parameters, and x_l^h denotes the activation of token x at head h and layer l. The intermediate outputs during training are: $\{(x_l^h, label)_i\}_{i=1}^N$, where N is the number of samples, and $label$ represents the category of the response corresponding to the input of the LLM.

By introducing the Attention-Augmented MLP, we dynamically adjust the weights of different channels in the MLP, enhancing important features while suppressing noise. The trained hallucination classification probes are used to extract intermediate layer representations of the large model and compute editing directions. The specific mathematical formula is:

$$\theta^h_{l,label} = \frac{1}{N}\sum_{i=1}^{N}(-1)^{\sigma\left(x^h_{l,label}\right)_i}, \quad h \in \text{Top Heads}. \tag{5}$$

Here, σ takes values from $\{0,1\}$, representing true or false Q+A pairs, respectively. The editing direction is computed based on the attention heads with the highest factual accuracy obtained by the probes.

During inference, the computed editing direction is incorporated into the process through the following formula:

$$x_{l+1} = x_l + \sum_{h=1}^{H} Q^h_l \left(\text{Att}^h_l\left(P^h_l x_l\right) + \alpha\sigma^h_l\theta^h_{l,label}\right) \tag{6}$$

where fine-grained hallucination direction intervention editing is achieved through the residual term, guiding the model's output toward factual correctness.

Sampling-based Dynamic Prompt Weight Adjustment

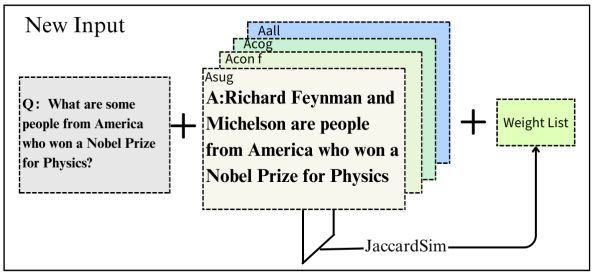

Figure 3: New input prompt and weights

In the second step of our overall framework, we adopted a psychology-oriented hallucination editing direction. Utilizing the obtained fine-grained editing directions, we performed parallel reasoning with the large model, generating four sampled outputs. These outputs included three responses based on single fine-grained editing directions and one response that did not distinguish between hallucination types. Subsequently, we employed the Jaccard formula(Bag, Kumar, & Tiwari, 2019) to dynamically calculate the weights for these three edited responses. The specific Jaccard calculation formula is as follows:

$$\text{Jaccard Sim}(A_i,A_j) = \frac{|\text{set}(A_i.split) \cap \text{set}(A_j.split)|}{|\text{set}(A_i.split) \cup \text{set}(A_j.split)|} \tag{7}$$

After obtaining the similarity calculation formula, we computed the overall similarity score for each response. The mathematical formula is as follows:

$$C_i = \frac{1}{2}\left(\text{sim}(A_i,A_j) + \text{sim}(A_i,A_k)\right) \tag{8}$$

Once the weights for each response were calculated, we compiled them into a weight list and concatenated the final input prompts. The specific formula is as follows:

$$A_{final} = \mathcal{M}(prompt \| Q \| A_{cog} \| A_{sug} \| A_{conf} \| A_{all} \| W) \tag{9}$$

Here, W represents the calculated weight list. In the PGFES scheme, we used A_{all} and the highest-weighted edited response as the primary references for the final response.

Experiments

Datasets

Our experimental datasets are derived from three publicly available datasets: TruthfulQA(Lin, Hilton, & Evans, 2022), FACTOR(Muhlgay et al., 2023), and Natural Questions(Kwiatkowski et al., 2019).

TruthfulQA is a dataset specifically designed to evaluate the propensity of language models to generate truthful responses to open-ended questions, aiming to detect and reduce the phenomenon of hallucination in model-generated content. The dataset spans multiple domains, including science, history, and health. Notably, the categories of our psychology-guided fine-grained hallucination classification do not conflict with the domains covered by the datasets. In our experiments, 50% of the TruthfulQA samples were used to validate the effectiveness of fine-grained hallucination editing methods.

The open-ended questions in this dataset are primarily evaluated based on two metrics: info and truth. The info metric measures the diversity of generated content, while the truth metric assesses the factual accuracy of the generated content. Given that the original GPT-3 model used for evaluation is no longer available, we employed the DeepSeek API (Liu et al., 2024) to assess two metrics in our experimental evaluation.

The FACTOR dataset comprises three subsets built from historical news articles, expert Q&A, and Wikipedia content, respectively. Natural Questions is a Google open-source dataset of real user queries. In our experiments, FACTOR's news and wiki subsets, along with Natural Questions, were utilized to evaluate the model's out-of-distribution generalization capability for effectively handling diverse real-world scenarios.

Baselines

We employed the llama2-7b-chat version as the base model and compared the following methods based on it:

- **Instruction Fine-tuning**: We utilized the Q&A texts from Wikipedia for instruction fine-tuning.
- **Contrastive Decoding**: We selected DoLa for experimentation. This method enhances the truthfulness of the

model's outputs by contrasting the logits outputs across different layers of the LLM.

- **Representation Editing**: The ITI method is utilized, which identifies the factual direction within attention heads and edits representations during the inference process to reduce the probability of the model generating hallucinated answers.
- **Sampling Optimization**: We primarily compared two methods, USC and ID. These methods enhance the self-consistency of the model's outputs by sampling multiple responses, thereby improving the truthfulness of the model's answers.

All experiments were conducted under the standard settings of TruthfulQA. The experimental results for contrastive decoding and sampling optimization were sourced from the original papers of DoLa and ID, respectively. The results for ITI were obtained by reproducing the experiments using its publicly available models and data.

Configuration

Our experiments were conducted on an NVIDIA A100 GPU with 80GB of memory. The fine-grained hallucination probes employs four-layer MLPs with dimensions [128 $\rightarrow$ 512, 512 $\rightarrow$ 64, 64 $\rightarrow$ 8, 8 $\rightarrow$ 1]. SE blocks were incorporated into the middle two layers to enhance feature representation. We utilized the AdamW optimizer with a learning rate of 1e-4 and trained the model for 10 epochs. Based on experimental analysis, we ultimately selected the editing strength α = 10 and the number of editing layers k = 12 for fine-grained hallucination editing. This configuration demonstrated strong applicability in open-ended text generation tasks.

Main Results on TruthfulQA

Table 1: Main results on TrutufulQA

Method	info	truth	I*T
llama2-7b-chat	0.9390	0.5902	0.5541
Instruction fine-tuning	0.9268	0.5951	0.5515
Dola	0.9439	0.5804	0.5478
USC	0.9608	0.5536	0.5319
ITI	0.9414	0.5975	0.5624
SR	0.9585	0.6170	0.5913
ID	0.9707	0.6073	0.5895
Ablation Experiment			
w/o Weight list	0.9512	0.6170	0.5869
w/o AA-MLP	0.9414	0.6073	0.5717
PGFES(I&T Balance)	0.9537	0.6268	0.5978
PGFES	**0.9292**	**0.6658**	**0.6187**

Table 1 compares the performance of PGFES with previous methods on the TruthfulQA dataset. PGFES achieved the best results in generative open-ended answering tasks, with an I*T (info*truth) score improvement of 11.67% over the baseline model Llama2-7b-chat(Touvron et al., 2023), particularly in the truth score. In contrast, instruction-tuning methods improved the truth score but reduced the info score, likely due to their negative impact on response diversity(Gekhman et al., 2024).

Compared to DoLa, PGFES improved truth metrics. DoLa's removal of shallow lexical information may reduce response richness(Fayyaz, Aghazadeh, Modarressi, Mohebbi, & Pilehvar, 2021), while PGFES optimizes the linear probes approach from ITI, achieving a 10.01% I*T score improvement. Specifically, PGFES introduces AA-MLP as a factual direction probe, more accurately identifying attention heads for factual generation. It also trains separate probes for different hallucination types, enabling fine-grained guidance and outperforming ITI in representation editing accuracy.

Against USC, PGFES showed a slight decline in info but improved the I*T metric. USC focuses on self-consistency, while PGFES maintains consistent inference temperature to preserve editing effectiveness. As an optimization of USC, PGFES achieved a 4.95% I*T score improvement over ID.

We enlisted three software engineering undergraduates who were not involved in data annotation to evaluate the truth scores. They compared the standard answers and the incorrect answers generated by LLMs and gave yes/no responses. The table 2 below shows the results of the human evaluation. The results show that the performance of PGFES is better than that of the other methods. We evaluated the inference efficiency of PGFES and prior sampling optimization methods. The results, as shown in the table 2, indicate that the time consumption of PGFES and USC is comparable and significantly lower than that of other methods. Although the proposed method introduces a certain degree of additional time consumption compared to the base model, this increase is considered acceptable in exchange for an improvement in the quality of inference results.

Furthermore, we conducted ablation studies, as evident from the table, where the removal of the weight list and the AA-MLP module respectively led to a decline in overall metrics, underscoring the efficacy of the integrated approach. Additionally, we have separately listed an experimental result that balances the scores of info and truth metrics. Although this result does not achieve the highest I*T score, it demonstrates outcomes under conditions of lower editing intensity and fewer layers.

Generalizability across more Benchmarks

To validate the out-of-distribution generalization capability of our proposed method for large language models, we conducted experiments on two out-of-distribution datasets: the FACTOR dataset and the Natural Questions dataset. In Table 3, the experimental results demonstrate that, compared to the baseline and ITI, PGFES achieves performance improvements on both datasets. Notably, on the Natural Questions dataset, PGFES shows a significant improvement of 5.88 points, which strongly supports the generalization ability of

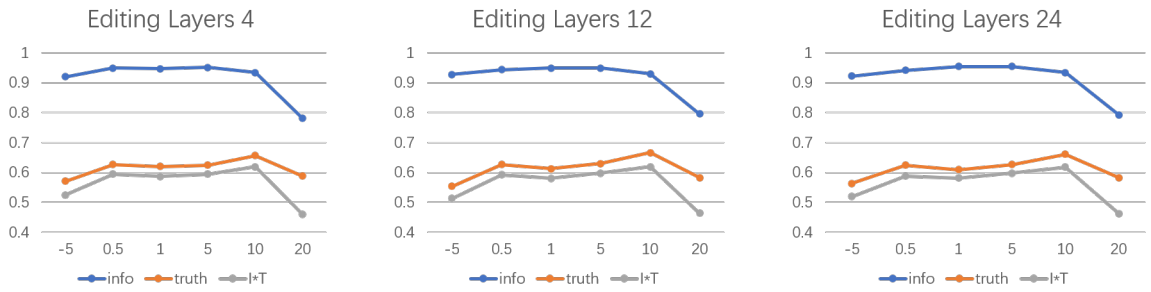

Figure 4: Trend graphs of info and truth scores for different editing strengths and layers

Table 2: Human evaluation and latency on TrutufulQA

Method	Human evaluation truth	Latency(ms/token)
llama2-7b-chat	0.5390	0.10
USC	0.5938	0.93
SR	0.6406	1.97
ID	0.6562	1.13
PGFES	**0.6755**	**0.96**

Table 3: Generalizability across more benchmarks

Method	News	Wiki	NQ
llama2-7b-chat	64.67	56.95	54.90
ITI	53.26	43.82	57.83
PGFES	**65.88**	**57.14**	**60.78**

our method in out-of-distribution scenarios.

Effect of Editing Layers and Strength

We conducted a series of experiments to analyze the impact of editing intensity and editing depth on the performance of LLMs in the PGFES. The experimental results indicate that, under the same editing depth, as the editing intensity increases from negative values to 20, the comprehensive I*T score of the model's output initially rises and then declines, reaching its peak at an editing intensity of 10. Further analysis of the info and truth scores reveals that, except for negative fine-tuning edits, the info score decreases as the editing intensity increases, while the truth score exhibits the opposite trend. This suggests that excessively high editing intensity may impair the diversity of information in the responses generated by the LLM.

Additionally, using the case of an editing intensity of 10 and an editing depth of 4 as a baseline, we recorded the info and truth scores as the editing depth increased. As can be seen from the figure 5, the overall trend shows a lower info score and a moderate increase in the truth score.

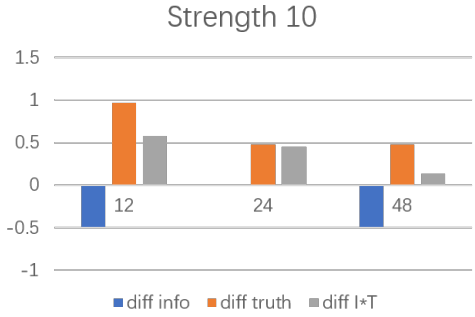

Figure 5: Trend graph of score changes with editing strength 10 and layers 4 as Baseline

Conclusion

This paper proposes a two-stage fine-grained optimization method (PGFES) to address hallucination issues in LLMs, integrating psychological theories and knowledge editing techniques to optimize the performance of LLMs. We introduce AA-MLPs for fine-grained direction extraction and employ a dynamic weight adjustment sampling strategy, significantly enhancing the authenticity of the output. The PGFES method has demonstrated significant efficacy on the open-source Llama 2 model. However, due to its reliance on editing during the inference process, its application to closed-source LLMs such as GPT-4 is currently not feasible. Future work will focus on optimizing this framework by incorporating more nuanced and in-depth concepts from cognitive science. A current limitation of the PGFES method is its dependence on manual data annotation, which may impede the scalability of the solution. Therefore, subsequent research will concentrate on developing an automated psychological hallucination classification model to replace manual intervention while ensuring data quality.

Acknowledgments

This work was supported by the National Natural Science Foundation of China(62471090), Sichuan Provincial Natural Science Foundation Project(23NSFSC0422), Central University Fund(ZYGX2024Z016).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmad, M. A., Yaramis, I., & Roy, T. D. (2023). Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*.

Asaad, G., & Shapiro, B. (1986). Hallucinations: theoretical and clinical overview. *The American journal of psychiatry*.

Bag, S., Kumar, S. K., & Tiwari, M. K. (2019). An efficient recommendation generation using relevant jaccard similarity. *Information Sciences*.

Banerjee, S., Agarwal, A., & Singla, S. (2024). Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*. doi: 10.48550/arXiv.2409.05746

Berberette, E., Hutchins, J., & Sadovnik, A. (2024). Redefining "hallucination" in llms: Towards a psychology-informed framework for mitigating misinformation. *arXiv preprint arXiv:2402.01769*.

Bouyamourn, A. (2023). Why llms hallucinate, and how to get (evidential) closure: Perceptual, intensional, and extensional learning for faithful natural language generation. *arXiv preprint arXiv:2310.15355*. doi: 10.48550/arXiv.2310.15355

Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2023). Discovering latent knowledge in language models without supervision. In *The eleventh international conference on learning representations.* Retrieved from https://openreview.net/forum?id=ETKGuby0hcs

Cao, B., Lin, H., Han, X., & Sun, L. (2024). The life cycle of knowledge in big language models: A survey. *Machine Intelligence Research*.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., & Edwards, H. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*. doi: 10.48550/arXiv.2107.03374

Chen, X., Aksitov, R., Alon, U., Ren, J., Xiao, K., Yin, P., ... Zhou, D. (2023). Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.

Cheng, Y., Liang, X., Gong, Y., Xiao, W., Wang, S., Zhang, Y., ... others (2024). Integrative decoding: Improve factuality via implicit self-consistency. *arXiv preprint arXiv:2410.01556*.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J. R., & He, P. (2024). Dola: Decoding by contrasting layers improves factuality in large language models. In *The twelfth international conference on learning representations.* Retrieved from https://openreview.net/forum?id=Th6NyL07na

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*. doi: 10.1093/jla/laae003

Fayyaz, M., Aghazadeh, E., Modarressi, A., Mohebbi, H., & Pilehvar, M. T. (2021). Not all models localize linguistic knowledge in the same place: A layer-wise probing on bertoids' representations. In *Proceedings of the fourth blackboxnlp workshop on analyzing and interpreting neural networks for nlp.* Association for Computational Linguistics. Retrieved from http://dx.doi.org/10.18653/v1/2021.blackboxnlp-1.29 doi: 10.18653/v1/2021.blackboxnlp-1.29

Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., & Herzig, J. (2024). *Does fine-tuning llms on new knowledge encourage hallucinations?*

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the ieee conference on computer vision and pattern recognition.*

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... Liu, T. (2025, January). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*. doi: 10.1145/3703155

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., ... others (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., ... Petrov, S. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Lai, E. R. (2011). Metacognition: A literature review.

Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P. N., Shoeybi, M., & Catanzaro, B. (2022). Factuality enhanced language models for open-ended text generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 34586–34599). Curran Associates, Inc.

Li, J., Yuan, Y., & Zhang, Z. (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*.

Li, K., Patel, O., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems.* Curran Associates, Inc.

Lin, S., Hilton, J., & Evans, O. (2022, May). TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers).* Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.229/ doi: 10.18653/v1/2022.acl-long.229

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... oth-

ers (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., … Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh conference on neural information processing systems.* Retrieved from https://openreview.net/forum?id=S37hOerQLB

Maiorca, V., Moschella, L., Norelli, A., Fumero, M., Locatello, F., & Rodolà, E. (2023). Latent space translation via semantic alignment. In *Thirty-seventh conference on neural information processing systems.* Retrieved from https://openreview.net/forum?id=pBa70rGHlr

Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., … Shoham, Y. (2023). *Generating benchmarks for factuality evaluation of language models.*

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., & Belinkov, Y. (2024). Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Rejeleene, R., Xu, X., & Talburt, J. (2024). Towards trustable language models: Investigating information quality of large language models. *arXiv preprint arXiv:2401.13086*.

Smith, A. L., Greaves, F., & Panch, T. (2023). Hallucination or confabulation? neuroanatomy as metaphor in large language models. *PLOS Digital Health*.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., … others (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*. doi: 10.48550/arXiv.2201.08239

Tonmoy, S., Zaman, S., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., … Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models.*

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *Elsevier*, 100211.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., … Shi, S. (2023). *Siren's song in the ai ocean: A survey on hallucination in large language models.*

Zheng, H., Xu, T., Sun, H., Pu, S., Chen, R., & Sun, L. (2024). Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. *arXiv preprint arXiv:2411.12591*.

Zhou, Y., Ni, Y., Gan, Y., Yin, Z., Liu, X., Zhang, J., … Chai, H. (2024). Are llms rational investors? a study on detecting and reducing the financial bias in llms. *arXiv preprint arXiv:2402.12713*.