International Journal of Advanced Engineering and Technology Research
ISSN: 3106-8677 (Print), ISSN: 3106-8685 (Online) | Vol. 1, No. 3, 2026

A Mathematical Investigation of Hallucination in Large Language Models

Linlin Su

University of Hong Kong, Hong Kong, China
leensu00@gmail.com

Abstract: This paper investigates the phenomenon of 'hallucinations' in large language models through a mathematical lens, analyzing their origins (including inadequate data and bias) and proposing three mitigation strategies: optimizing the reward function in reinforcement learning from human feedback (RLHF), employing low-probability tokens to enhance decoding strategies, and implementing uncertainty-based detection methods (such as SelfCheck-GPT). The study seeks to improve the precision and dependability of model results.

Keywords: LLMs, Hallucination, RLHF, Gaussian Process, Uncertainty Estimation, Decoding Strategy.

1. Introduction

In large language models (LLMs), hallucinations occur when the model produces irrelevant, illogical, or erroneous outputs. Definitions of hallucinations, distinctions between types (such as factual errors, nonsensical answers, or irrelevant information), and settings under which they occur are the main areas of research in this field. The goal of the mathematical study of hallucinations in LLMs, such as GPT-4 is to comprehend, identify, and minimize the situations in which these models generate outputs that are erroneous, incomprehensible, or unconnected. This community has been focusing on and categorizing and characterizing hallucinations, such as factual errors, illogical responses, or irrelevant information. The potential reasons leading to these hallucinations frequently ascribed to insufficient datasets, overfitting of models, or biases in the datasets (Zhang et al., 2023) [1]. Researchers propose mathematical models to comprehend and forecast hallucinations in LLMs. These models leverage computational theories, probability theory, and statistical analysis to explain how complicated interactions between different model components result in hallucinations (Lee, 2023) [2].

Additionally, this field of research also prioritizes the techniques for the identification, measurement, and reduction of hallucinogenic effects (Ji et al., 2023) [3]. It entails employing anomaly detection and natural language understanding to create benchmarks and metrics for gauging the frequency and intensity of hallucinations. Also, mitigation techniques include anything from adding external knowledge bases and post-processing filters to improving training data and changing model topologies. Meanwhile, GPT's self-supervised learning paradigm (Manakul et al., 2023) [4], which makes it more robust to learn from vast volumes of unlabeled data, is what has allowed for its accomplishments. The objective of the training procedure entails forecasting the next token in a series, which is to maximize the likelihood of creating the observed data given the model's parameters, is presented as a maximum likelihood estimation issue.

However, such Hallucination warrants a more rigorous extensive studies, especially for the Mitigation of LLM Hallucination. I yearn to investigate whether mathematical model architectural improvements would reduce hallucinations. In this proposal, three aspects of Mitigation are involved: Mitigation during reinforcement learning from human feedback (RLHF), Mitigation during RLHF, and exploiting uncertainty. We define Mitigation during RLHF in terms of training a reward function and improving the quality of training data for the reward model. We propose the low token-level output probabilities to design decoding strategies to mitigate hallucinations in LLMs during inference. Moreover, implementing a non-trivial challenge with SELFCHECK-GPT to assess the consistency with which different responses correspond to the estimate of uncertainty.

I further characterize the potential future research of evaluating, tracing, and eliminating hallucinations within LLMs, where we can progressively test and modify if the mathematical models can help mitigate Hallucination.

2. Literature Review on Mathematical Analysis of Hallucination

Fernandes, Madaan and Emmy (2023) [5] highlighted that the crucial role of human feedback in refining natural language generation by large language models, emphasizing that such models can generate problematic content. Hence, they will put more efforts on the enhancement of the supervised fine-tuned (SFT) LLMs by RLHF including developing a reward model (RW) to serve as a stand-in for human preference and maximizing the SFT model using feedback from the reward model (Schulman, 2023) [6]. Zhang et al. (2023) [1] argued that Hallucinations may result from the process of behavior copying during the SFT stage. And the honesty oriented RLHF approach can lead to over-conservatism in language models, compromising their helpfulness by making them overly cautious in providing information. Given this, Schulman (2023) [6] suggested resolving this issue during RLHF, which creating a unique reward function specifically to lessen delusions.

Although the special design of reward function with some statistical distributions might eliminating the hallucination, reducing hallucinations throughout the inference period might be more manageable and economical (Zhang et al., 2023) [1]. Huang and Chang (2023) [7] employed the idea that the next-token distribution is determined by contrasting the differences in logits from later versus earlier transformer

layers, utilizing the fact that factual knowledge in large language models is typically concentrated in specific layers. However, these approaches require accessing token-level output probabilities and suggests a shift towards exploring within a stricter black-box setting (Zhang et al., 2023) [1].

Moreover, Manakul, Liusie and JF Gales (2023) [4] proposed that "SelfCheck-GPT" is a sampling-based approach that fact-checks responses from black-box models, operating on the principle that responses containing hallucinated facts will tend to be inconsistent or contradictory. To further research on Consistency-based estimation methods, Liu et al. (2021) [8] found that defining new prompting functions, the model can perform few-shot or even zero-shot learning, adapting to new tasks with minimal or no labeled data which can be considered to address the challenges of leveraging uncertainty in Hallucination.

3. Prospective Methodologies

In the context of large-scale language models (LLM), mathematical reasoning poses a significant challenge, and the extent to which it correlates with LLM capacity has yet to be thoroughly investigated (Bıyık et al., 2023) [9]. As a result, we plan to initiate a mathematical investigation into the phenomenon of hallucinations within a generative pretrained transformer (GPT) model (Lee, 2023) [2]. Using concepts from probability theory and information theory, we rigorously define and quantify hallucination and creativity in order to determine the optimal equilibrium that maximizes model performance across a variety of tasks.

3.1. Integrating Gaussian Process into Reward Function

To propose an approach to learning based on preferences, in which human feedback is provided through trajectory comparisons (Bıyık et al., 2023) [9]. We do not impose stringent limitations on the reward function's structures. Alternatively, we utilise a Gaussian process to represent the reward function and provide a mathematical approach to dynamically adjust the model based on human preferences.

We first study the gaussian process that defined by its mean function m(x) and covariance function k (x, x′), where x and x′ are points in the input space. Then, assuming the true reward function R (s, a) for state s and action a is unknown and needs to be estimated. We model R (s, a) as a sample from a GP: R (s, a) ~ GP (m (s, a), k ((s, a), (s′, a′)). In this paper, we aim to design with the inspiration of decision-making in autonomous driving, focus on maximum seeking of unknown environment fields.

The reward function design used as a replacement for DNN model, denoted by W (s, a; θ_w) which is parameterized by parameters θ_w (Lin et al., 2022) [10]. Then assuming the N training data points, x = (x_1, ..., x_N), y = (y_1, ..., y_N)T needs to satisfy

$$f(\cdot) \sim GP(0,\ K(\cdot,\cdot)),\ y = f(x) + \epsilon,\ \epsilon \sim N(0, \sigma_n^2 I) .$$

We consider the unknown objective function f(x) is located at x with a noise, the objective observations D={(x_i,y_i)|i=1, …, t), and Kernel fucntions K(·, ·) (Skalse et al., 2022) [12].

After that, we add the Bayesian principle to Gaussian likelihood Condition, we obtain the mean and variance of the output y_* by the input x_*:

$$Mean:\ K(x_*, x)[(K(x,x) + \sigma_n^2 I)]^{-1} y$$

$$Variance:\ K(x_*, x_*) + \sigma_n^2 I - K(x_*, x)[(K(x,x) + \sigma_n^2 I)]^{-1} K(x, x_*)$$ (Lin et al., 2022) [10].

From this, in Gaussian Process (GP) regression, the negative log marginal likelihood (NLML) is often used as the loss function to optimize the hyperparameters of the model. The NLML provides a measure of how well the model explains the observed data under the current set of hyperparameters. By minimizing the NLML, we find the hyperparameters that make the observed data most probable, thereby improving the predictive performance of the GP model. Ideally, the model W (s, a; θ_w) can be trained to mitigate the hallucination after giving a reasonable state reward by using marginal likelihood (where M is the number of tasks, MN represents matrix). Therefore, we get an NLML function and maximize the multiple tasks:

$$p(y \mid X) \sim N(0, O(M^3 N^3))$$

$$-log\ p\ (y \mid x) = \frac{N}{2} log(2\pi) + \frac{1}{2} log|\ K(x,x) + \sigma_n^2 I)\ | \ + \frac{1}{2} y^T (K(x,x) + \sigma_n^2 I)^{-1} y$$

$$-L_{NLL} = \frac{MN}{2} log(2\pi) + \frac{1}{2} log|\ K(x,x) + \sigma_n^2 I)\ | \ + \frac{1}{2} y^T (K(x,x) + \sigma_n^2 I)^{-1} y$$ (Long et al., 2023) [11]

Given the intricate surface of the objective function and the Gaussian system, one potential problem is ensuring that the solution is close to the optimum. Re-estimating using the new numbers as the starting point and shocking the estimations by a set number of standard deviations is one way to solve the problem. The different estimations' solutions will probably converge to the global optimum if they tend to be comparable. The simulated annealing approach, a well-liked machine learning optimization technique, incorporates this concept and can be used to test the stability of our parameter estimations.

3.2. Employing Low-Probability Tokens to Decoding Strategies

The probability of producing low-probability tokens that differ from the predicted output increases as the context gets less instructive and the gap between the highest and subsequent probabilities shrinks. This finding emphasizes the importance of recognizing that confusing input circumstances might create hallucinations even in well-trained GPT models (Lee, 2023) [2].

We first assume a joint distribution P(x)= $\prod_{i=1}^{n} p(x_i | x_1, x_2, \ldots, x_{i-1})$ of the sequences of tokens X={$x_1, x_2, \ldots, x_n$}, where the conditional probabilities p($x_{i+1}|x_1, x_2, \ldots, x_{i-1}; \Theta$} are calculated for all tokens and for all next potential places i

$\in 1, 2, \ldots, n$.

Then we try to generate a token x_{i+1} with low-probability, and then predict the following tokens x_{i+2} based on past tokens $x_1, x_2, \ldots, x_i$, so we applied the probability distribution to the hallucination phenomenon, the n+1 prediction is defined as bellow:

$$H(x_{i+1}|x_1, x_2, \ldots, x_i; \Theta) = -\sum_{x'} p(x_{i+1}|x_1, x_2, \ldots, x_i; \Theta) log p(x_{i+1}|x_1, x_2, \ldots, x_i; \Theta)$$ (Lee, 2023) [2].

Where $p(x_{i+1}|x_1, x_2, \ldots, x_i; \Theta) = \frac{exp(z_{i+1}(x_{i+1;}\Theta))}{\sum_{x'} exp(z_{i+1}(x'\Theta))}$ the x' represents the next position i + 1. This approach has an uncertainty which is increased when GPT models generate tokens with low probability. Thus, the estimated probability $p(x_{i+1})$ are distributed in a way that the gap between the highest and subsequent probabilities is very minimal when the input context does not give enough information for a clear and optimal token choice.

In the future, we might correlate higher levels of uncertainty in GPT models to the development of low-probability tokens to eliminate hallucination.

3.3. SELFCHECK-GPT with Prompt Functions

Scheck and Schütze (2021a, b) [13] suggested using cloze-style exercises to fine-tune pre-trained language models (PLMs) as a prompting strategy to bridge the gap between the pre-training and fine-tuning stages. A verbalizer ν and a preset template T are the two main parts of a prompt, according to formal definitions. Wrapped with the template is each input sample x. In this paper, we would like to consider the uncertainty estimation with prompts to mitigate the hallucinations.

We will introduce sample S^n samples are converted to pseudo labels y_i^n by support set with highest $p(y_i|x)$ for class i,

$$Let \qquad \hat{y}_i = (\frac{p(y_i|x)}{p(\nu(y_i)}) / (\sum_{j=1}^{C} \frac{p(y_i|x)}{p(\nu(y_i)}) \; and \; S_{promot}(i) = \frac{1}{N} \sum_{n=1}^{N} y_i^n .$$

(Manakul et al., 2023 and Ye et al., 2023) [4] [14]

Then, calculating uncertainty for each sample x, we obtained

$$u(x) = -\sum_{i=1}^{C} \hat{y}_i log \; \hat{y}_i$$. (Manakul et al., 2023) [4]

While prompt-based approaches are anticipated to yield superior performance when prompt functionalities are incorporated into PLMs, this comes at the expense of increased sample quantity. Hence, we may further examine the performance by varying the amount of samples to observe the eradication of hallucinations.

4. Implications and Contributions

An examination of hallucinations in huge language models using mathematical analysis signifies a significant progress in the domain of artificial intelligence, tackling a crucial obstacle in guaranteeing the precision and dependability of these systems. Through the analysis of the root causes of incorrect outputs, this research provides valuable insights that are crucial for improving the design of models and training methods. Consequently, this improves the reliability of AI applications in several critical fields, such as healthcare and legal consultation. Furthermore, this investigation contributes to more extensive deliberations in the field of AI ethics and safety, guaranteeing that these potent instruments conform to human values and function inside secure boundaries. The results of this inquiry not only advance AI technology but also offer crucial interdisciplinary insights that could potentially impact subjects such as cognitive science and neuroscience.

References

[1] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023, September 24). Siren's song in the AI Ocean: A survey on hallucination in large language models. arXiv.org. https://arxiv.org/abs/2309.01219

[2] Lee, M. (2023). A mathematical investigation of hallucination and creativity in GPT models. Mathematics, 11(10), 2320. https://doi.org/10.3390/math11102320

[3] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

[4] Manakul, P., Liusie, A., & Gales, M. J. F. (2023, October 11). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. arXiv.org. https://arxiv.org/abs/2303.08896

[5] Fernandes Fernandes, P., Madaan, A., Liu, E., Farinhas, A., Martins, P. H., Bertsch, A., de Souza, J. G. C., Zhou, S., Wu, T., Neubig, G., & Martins, A. F. T. (2023, June 1). Bridging the gap: A survey on integrating (human) feedback for natural language generation. arXiv.org. https://arxiv.org/abs/2305.00955

[6] YouTube. (2023, April 20). John Schulman - reinforcement learning from human feedback: Progress and challenges. YouTube.

[7] Huang, J., & Chang, K. C.-C. (2023, May 26). Towards reasoning in large language models: A survey. arXiv.org. https://arxiv.org/abs/2212.10403

[8] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021, July 28). Pre-train, prompt, and predict: A systematic survey of prompting methods in Natural Language Processing. arXiv.org. https://arxiv.org/abs/2107.13586

[9] Bıyık, E., Huynh, N., Kochenderfer, M. J., & Sadigh, D. (2023). Active preference-based Gaussian process regression for reward learning and optimization. The International Journal of Robotics Research. https://doi.org/10.1177/02783649231208729

[10] Lin, J., Fried, D., Klein, D., & Dragan, A. (2022, April 5). Inferring rewards from language in context. arXiv.org. https://arxiv.org/abs/2204.02515

[11] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022, March 4). Training language models to follow instructions with human feedback. arXiv.org. https://arxiv.org/abs/2203.02155

[12] Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022, September 27). Defining and characterizing reward hacking. arXiv.org. https://arxiv.org/abs/2209.13085

[13] Schick, T., & Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and Natural Language Inference. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. https://doi.org/10.18653/v1/2021.eacl-main.20

[14] Ye, H., Liu, T., Zhang, A., Hua, W., & Jia, W. (2023, September 13). Cognitive mirage: A review of hallucinations in large language models. arXiv.org. https://arxiv.org/abs/2309.06794