International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677● Website: www.ijsat.org● Email: editor@ijsat.org

Does Seed Matter? Investigating the Effect of Random Seeds on LLM Accuracy

Praneeth Vadlapati

University of Arizona
praneethv@arizona.edu
ORCID: 0009-0006-2592-2564

Abstract

The recent rapid advancements in Natural Language Processing (NLP) include Large Language Models (LLMs), which have transformed the landscape of NLP. However, a challenge exists in understanding the factors that affect model accuracy and consistency. Randomseed values are frequently used to maintain consistency in the responses. This research investigates the impact of the usage of random seeds on the accuracy of LLM responses to challenging questions. The study investigates the potential for seed-specific hallucination patterns across various types of questions. The findings of the experiment have demonstrated a considerable amount of variability across multiple questions using different seeds. The experiment discovers a new factor that affects the reliability of LLMs. The discovery is crucial for mission-critical applications that use LLMs since accuracy and consistency are of high importance.The source code is available atgithub.com/Pro-GenAI/PromptSeed.

Keywords: Large Language Models (LLMs), hallucinations, prompt engineering, random seed, accuracy, Natural Language Processing (NLP)

I. Introduction

Large Language Models (LLMs) based on Transformer architecture [1] are recent indispensable advancements in Artificial Intelligence (AI) that demonstrated remarkable capabilities in generating text content and in answering questions based on context [2], [3], [4], [5]. Despite these advancements, challenges persist, particularly in addressing hallucinations and accuracy inconsistencies across multiple responses [6], [7], [8], [9]. The concept of controlling randomness and allowing reproducibility through seed values plays a crucial role in influencing the deterministic nature of the model responses [10], [11], [12], [13]. While randomness enhances the creativity of responses, the role of randomness in altering model accuracy remains underexplored. Exploring and addressing these issues is essential for the utilization of LLMs in sensitive or mission-critical applications, where consistency and accuracy are essential.

A. Proposed experiment

This research explores the role of random seeds in shaping LLM responses to various computational tasks and compares them with the cases of no usage of random seeds. The experiment employs a systemic approach to determine whether the selection of seed values influences the consistency, accuracy, and likelihood of hallucinations in LLM responses. The findings aim to provide an understanding of the effects of multiple values for random seeds to facilitate the process of improvements in model accuracy and the decisions of their deployment in critical applications.

B. Related work

Ji et al. (2023) [14] discuss hallucinations as a prevalent issue in the Natural Language Generation (NLG) and categorize them into intrinsic and extrinsic types based on their origins and effects. Their work does not address the role of randomness introduced by seeds. Prompt is another impactful aspect of LLM behavior. Prompt engineering has been shown to affect response quality significantly. Reynolds and McDonell (2021) [15] emphasize that minor variations in prompts have the potential to cause a substantial difference in the LLM responses. A considerable amount of work exists on LLM hallucinations. However, the existing work focuses on hallucination and variability in responses and does not explore the interaction between prompt engineering and settings such as seed values. However, existing research has not explored the impact of settings such as seed values on LLM accuracy and hallucinations. This study is designed to investigate the impact of controlling randomness using seed values on the accuracy and hallucinations of LLMs. By varying the seed values and analyzing their impact on the responses, this research provides a novel perspective on factors that influence the reliability of LLMs.

II. Methods

A. Selecting and loading an LLM

The experiments require an LLM with demonstrated performance and accuracy scores to ensure reliable response generation to answer regular queries without necessarily utilizing a random seed. The results are required to be meaningful, relevant, and consistent. Hence, GPT-3.5 [16] is selected as the model by considering its demonstrated high scores in accuracy and performance across a variety of NLP tasks. The model is accessed through the API, which allows custom parameters such as seed value, enabling a comprehensive exploration of randomness in response generation and accuracy.

B. Preparation of questions

Four questions are designed manually to include computational and linguistic tasks. These were created to test the logical reasoning and language processing capabilities of the LLM. These tasks represent different levels of complexity, which allow us to evaluate the influence of random seed values across diverse problem types. The expected correct answers are defined alongside the questions to facilitate the automated evaluation process. The questions and the expected answers are mentioned in the table below.

TABLE : 1 QUESTIONS CREATED FOR THE EXPERIMENT

Index	Question name	Question	Expected answer
1	2^4	Answer 2^4 and write the answer inside backticks. Example: `123`.	16
2	30-24	Remove 24 from double of 15 and answer inside backticks. Example: `123`.	6
3	Vowel count	Find number of vowels in “aeronautics” and answer inside backticks. Example:	6

Index	Question name	Question	Expected answer
		`123`.	
4	Text reverse	Reverse the text “strawberry” and write the reverse in backticks. Example: `apple`. Do not write code.	yrrebwarts

C. Selecting random seed values

The experiment utilized a range of commonly used random seed values. The values are designed to use the seed values commonly used in real-world AI-based applications to ensure that the experiment resembles real-world scenarios, which ensures that the findings are relevant to practical scenarios. A diverse set of seeds are included in the experiment to capture a diverse range of randomness effects. The seed values are mentioned in the table below.

TABLE: 2 SEED VALUES CREATED FOR THE EXPERIMENT

Index	Seed
1	None
2	0
3	10
4	20
5	64

D. Testing using LLM

The created prompts and random seed values are utilized with the LLM to generate responses. Ten attempts are conducted for each test case to ensure accuracy in the process of evaluating accuracy. Answers generated across all attempts are extracted from the responses and compared with predefined correct answers using an automated evaluation to compute accuracy. Accuracy calculated at each test case is the metric used for the evaluation of the responses.

III. RESULTS

A. Accuracy of responses at various seed values

The accuracy calculated across multiple attempts in each test case is mentioned in the table below.

TABLE: 3 ACCURACY OF RESPONSES

Index	Question name	Accuracy at different seeds				
		None	*0*	*10*	*20*	*64*
1	2^4	100%	100%	100%	100%	100%
2	30-24	80%	80%	100%	**0%**	100%
3	Vowel count	**10%**	**30%**	**0%**	**0%**	100%

Index	**Question name**	**Accuracy at different seeds**				
		None	***0***	***10***	***20***	***64***
4	Text reverse	100%	100%	100%	100%	100%

IV. Discussion

Despite consistent results for multiple seed values across multiple queries, the findings reveal the significant impact of random seed values on LLM performance. The most common mistake made by the model was the generation of incorrect answers. While randomseeding aims to standardize responses, the choice of seed has the potential to impact variability. This finding has a considerable discovery regarding the risks of using seeding for reproducibility in LLM responses. The identification of seed-specific hallucination patterns creates questions about the underlying mechanisms driving these phenomena. The experiment leaves a gap in the exploration of other parameters passed along with the prompt that alter the accuracy and reliability of LLMs, especially in critical applications. The findings demonstrate that the implementation of LLMs in critical applications with custom seeds necessitates careful selection of the seed and a rigorous testing process to ensure reliability before implementation in critical applications in the real world.

V. Conclusion

This experiment highlights the critical role of random seeds in shaping the accuracy and reliability of LLM responses. By systematically varying seed values, the study demonstrates that randomness, which enhances creativity, is a potential source of hallucinations and variability in responses. The findings require significant consideration in the selection and evaluation of seed values in LLM deployment. The findings introduce new considerations regarding factors influencing the accuracy of LLM responses. Future work should investigate the impact of other hyperparameters, such as temperature, on LLM responses. It can also combine multiple settings, such as temperature and seed. Future research should test LLM using multiple benchmarks with different values in hyperparameters. Addressing the hallucination concerns discovered in this paper is crucial for the development of reliable user-centered AI-based systems in the future.

REFERENCES

[1] A. Vaswani et al., “Attention is all you need,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS’17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 6000–6010. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[2] J. Robinson and D. Wingate, “Leveraging Large Language Models for Multiple Choice Question Answering,” in The Eleventh International Conference on Learning Representations, Feb. 2023. [Online]. Available: https://openreview.net/forum?id=yKbprarjc5B

[3] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large Language Models are Zero-Shot Reasoners,” Jan. 2023, arXiv:2205.11916. [Online]. Available: https://arxiv.org/abs/2205.11916

[4] E. Kamalloo, N. Dziri, C. L. A. Clarke, and D. Rafiei, “Evaluating Open-Domain Question Answering in the Era of Large Language Models,” May 2023, arXiv:2305.06984. [Online]. Available: http://arxiv.org/abs/2305.06984

[5] K. Singhal et al., “Towards Expert-Level Medical Question Answering with Large Language Models,” May 2023, arXiv:2305.09617. [Online]. Available: http://arxiv.org/abs/2305.09617

[6] R. Azamfirei, S. R. Kudchadkar, and J. Fackler, “Large language models and the perils of their hallucinations.,” Crit Care, vol. 27, no. 1, p. 120, Mar. 2023, doi: 10.1186/s13054-023-04393-x.

[7] N. Mündler, J. He, S. Jenko, and M. Vechev, “Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation,” May 2023, arXiv:2305.15852. [Online]. Available: http://arxiv.org/abs/2305.15852

[8] N. Dziri, S. Milton, M. Yu, O. Zaiane, and S. Reddy, “On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?,” in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 5271–5285. doi: 10.18653/v1/2022.naacl-main.387.

[9] N. McKenna, T. Li, L. Cheng, M. J. Hosseini, M. Johnson, and M. Steedman, “Sources of Hallucination by Large Language Models on Inference Tasks,” May 2023, arXiv:2305.14552. [Online]. Available: http://arxiv.org/abs/2305.14552

[10] S. Qian et al., “Are My Deep Learning Systems Fair? An Empirical Study of Fixed-Seed Training,” in Advances in Neural Information Processing Systems, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., Nov. 2021. [Online]. Available: https://openreview.net/forum?id=kLWGdQYsmC5

[11] P. Madhyastha and R. Jain, “On Model Stability as a Function of Random Seed,” in Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), M. Bansal and A. Villavicencio, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 929–939. doi: 10.18653/v1/K19-1087.

[12] S. Raste, R. Singh, J. Vaughan, and V. Nair, “Quantifying Inherent Randomness in Machine Learning Algorithms,” SSRN Electronic Journal, Jun. 2022, doi: 10.2139/ssrn.4146989.

[13] S. Dutta, A. Arunachalam, and S. Misailovic, “To Seed or Not to Seed? An Empirical Analysis of Usage of Seeds for Testing in Machine Learning Projects,” in 2022 IEEE Conference on Software Testing, Verification and Validation (ICST), Los Alamitos, CA, USA: IEEE Computer Society, Apr. 2022, pp. 151–161. doi: 10.1109/ICST53961.2022.00026.

[14] Z. Ji et al., “Survey of Hallucination in Natural Language Generation,” ACM Comput. Surv., vol. 55, no. 12, Mar. 2023, doi: 10.1145/3571730.

[15] L. Reynolds and K. McDonell, “Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm,” in Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, in CHI EA ’21. New York, NY, USA: Association for Computing Machinery, May 2021. doi: 10.1145/3411763.3451760.

[16] “GPT-3.5 Turbo (gpt-3.5-turbo-0301) [Language Model],” OpenAI. Accessed: May 27, 2023. [Online]. Available: https://openai.com/index/introducing-chatgpt-and-whisper-apis/