From Model Multiplicity to Prompt Multiplicity: Emerging Arbitrariness Concerns in the Age of Generative AI

Prakhar Ganesh
McGill University
Mila - Quebec AI Institute
prakhar.ganesh@mila.quebec

Motivation

Model multiplicity is the existence of multiple models that perform similarly well on a specific task, despite different underlying decision boundaries (Black, Raghavan, and Barocas 2022). If managed well, model multiplicity gives us the freedom to select better models from a set of 'good models' in ways that minimize harm (Black, Raghavan, and Barocas 2022; Ganesh et al. 2023; Black et al. 2024). However, model multiplicity also marks the unavoidable presence of arbitrariness in model selection that can impact individuals, necessitating a broader discussion on the expectations of AI decision makers in our society (Jain, Creel, and Wilson; Creel and Hellman 2022; Gomez et al. 2024; Kulynych et al. 2023).

The recent adoption of generative AI (GenAI) systems has fundamentally reshaped how developers design and deploy learning models (Yuan 2023). Traditionally, machine learning workflows centered around the training of task-specific models. In contrast, we are now seeing an increasing reliance on powerful pre-trained models, such as large language models (LLMs), with a growing focus on designing effective prompt structures to elicit desirable behavior (Dhar 2024; Amatriain 2024). In this new paradigm, prompts—not models—have become the focus of system design and control.

This has given rise to a new challenge: *prompt multiplicity*. Here, a prompt refers to both the textual instructions and other design choices when interacting with the GenAI system. Different prompts can produce significantly different outputs (Sclar et al. 2023; Voronov, Wolf, and Ryabinin 2024), and such differences persist even when the overall task performance is stable (Ganesh, Shokri, and Farnadi 2025). The shift from model-centric to prompt-centric interactions recontextualizes the problems of multiplicity. Thus, adapting the discussions of model multiplicity and extending them to novel concerns in GenAI, i.e., the study of prompt multiplicity, is critical to ensuring transparency and accountability in the emerging GenAI landscape.

Scope Traditional concerns around model multiplicity persist and are even exacerbated in the era of generative models (e.g., fear of a generative monoculture (Wu, Black, and Chandrasekaran 2024)). However, this extended abstract does not focus on the study of model multiplicity in GenAI systems. Instead, it centers on a different but related issue: prompt multiplicity. Specifically, it explores the challenges, implications, and opportunities that arise when diverse prompts are used with a single, fixed generative model.

Background

Model Multiplicity Multiplicity, as defined by Marx, Calmon, and Ustun (2020), is the occurrence of conflicting predictions across competing models. While the phenomenon of multiplicity in prediction problems is not new, recent research under the umbrella of *model multiplicity* (Black, Raghavan, and Barocas 2022) has broadened the discourse to its real-world impact and potential in advancing responsible AI practices. This includes work on leveraging multiplicity to select better models (Black, Raghavan, and Barocas 2022; Ganesh 2024), studying the trends of multiplicity with fairness and privacy (Kulynych et al. 2023; Long et al. 2023), better quantification and efficient assessment of associated risks (Watson-Daniels et al. 2023; Watson-Daniels, Parkes, and Ustun 2023; Hsu et al. 2024; Hsu and Calmon 2022; Ganesh 2024), and raising ethical and legal concerns on the implications of multiplicity within existing AI ecosystems (Black, Raghavan, and Barocas 2022; Creel and Hellman 2022; Black et al. 2024).

Prompt Sensitivity Prompt sensitivity has been extensively studied in the literature, revealing that even minor changes to the input can impact model behaviour (Lu et al. 2022; Shi et al. 2023; Sclar et al. 2023; Voronov, Wolf, and Ryabinin 2024). For instance, Lu et al. (2022) showed that simply shuffling the demonstrations in the prompt can affect accuracy, while Shi et al. (2023) showed that even irrelevant text in the prompt can change the output. Similarly, studies on prompt templates reveal that even adjustments like adding a space (Sclar et al. 2023; Voronov, Wolf, and Ryabinin 2024) can disrupt evaluations.

Methodology

A fundamental question in the age of GenAI is: *can we interpret the decisions of large-scale generative models?* Despite numerous attempts, our understanding of the internal workings of these systems remains limited and fragmented.

Interestingly, prompt multiplicity offers a promising entry point into this challenge. By examining how a generative model responds to different prompt structures, we can uncover patterns in its behavior. Thus, prompt multiplicity not only raises important questions about stability, but also provides a lens to better understand these complex systems. *We focus on prompt multiplicity, both as a phenomenon that raises critical issues for responsible deployment and as a methodological tool for probing GenAI models.*

From Prompt Sensitivity to Prompt Multiplicity As discussed, prompt sensitivity of GenAI systems has been widely recognized. However, prompt sensitivity literature tends to focus solely on accuracy, i.e., it treats prompts as levers for maximizing task performance (Sclar et al. 2023; Voronov, Wolf, and Ryabinin 2024), analogous to finding more accurate models in a model-centric perspective. These studies often overlook subtler but critical dimensions of model behavior, i.e., the stability of individual decisions.

Prompt multiplicity builds on this by shifting the focus from aggregate performance metrics to the variability in individual queries under different prompt structures (Ganesh, Shokri, and Farnadi 2025). Unlike prompt sensitivity, which is mainly concerned with finding prompts that yield higher accuracy, prompt multiplicity examines cases where prompt structures with similar accuracy levels create significantly different decision boundaries. This inconsistency challenges the assumption that a stable accuracy score indicates robust model behavior and reveals a deeper layer of unpredictability that can have real-world consequences.

In our recent study on hallucination evaluation in LLMs (Ganesh, Shokri, and Farnadi 2025), we found significant multiplicity (over 50% inconsistency in several benchmarks), despite stable accuracy reported by prior work. These findings underscore the risks of ignoring prompt multiplicity: it can mask the presence of hallucination-related harms and blur the distinction between random inconsistencies and consistently incorrect outputs. Extending this analysis to other domains, especially those considered immune to prompt sensitivity, could reveal hidden trends and enhance our understanding of GenAI systems.

Larger Attack Surface under Multiplicity Given the widespread use of GenAI models through APIs, restricting access via traditional means like rate limits has become impractical. Thus, a key vulnerability emerges due to prompt multiplicity—the ability of adversaries to test millions of prompts to uncover model behaviors not apparent through limited interaction. This enables actors to reverse-engineer or jailbreak models, elicit restricted outputs, and extract training data at an unprecedented scale (More, Ganesh, and Farnadi 2024; Nasr et al. 2023). In effect, prompt multiplicity vastly expands the attack surface, exposing brittleness and a lack of robustness of current GenAI systems.

Our recent work showed that real-world adversaries can exploit prompt multiplicity to effectively double the risk of data extraction (More, Ganesh, and Farnadi 2024). Through various case studies, including extracting pretraining data, detecting copyright violations, and retrieving personally identifiable information, we showed how even a naive adversarial strategy of exploiting prompt multiplicity can outperform existing extraction techniques. These findings underscore the need for more advanced defenses and a broader conversation around the implications of prompt multiplicity on the security of GenAI systems.

Auditing under Multiplicity In traditional model-centric settings, models often reflect design choices that may be influenced by arbitrary factors such as training data, architecture, or hyperparameters. However, in prompt-centric settings, users directly interact with the system via language prompts, and thus, the role of the end user becomes central.

Commercial platforms often design systems for a 'typical user' and avoid accountability for behaviors resulting from adversarial use (Aerni et al. 2025). While it is important to critique this limited responsibility, there are contexts in which the notion of a typical user is defensible. For instance, if an adversarial prompt is highly unlikely, the generated bias or toxicity resulting from it may not be a realistic risk. Thus, a nuanced approach is needed: one that acknowledges that some threat scenarios are outside the practical bounds of user behavior.

To address this, our framework AUGMENT (Automated User-Grounded Modeling and Evaluation of Natural language Transformations) (Chataigner et al. 2025) offers a systematic way to assess prompt multiplicity grounded in real user behavior. Rather than randomly generating prompt variations, AUGMENT produces controlled and demographically relevant paraphrases that retain semantic meaning and follow linguistic and stylistic norms. This approach simulates how actual users are likely to vary their prompts, leading to more accurate and representative audits. Our results highlight the need for more representative and structured approaches to prompt multiplicity in LLM auditing.

Quantifying Multiplicity in Language To adequately address the concerns of prompt multiplicity in GenAI, a robust framework for quantifying multiplicity is needed. Most existing discussions focus on structured MCQA formats, with limited scope. However, real-world interactions occur through natural language or other unstructured modalities, where responses can vary widely, making it far more challenging to define and detect multiplicity.

In unstructured settings, measuring consistency across diverse generations becomes particularly complex. While several metrics, such as semantic similarity for text (Farquhar et al. 2024) or image similarity (Song, Liu, and Shou 2024), have been proposed, these approaches depend heavily on third-party evaluators like LLM judges. These evaluators can introduce their own biases and inconsistencies, undermining their reliability (Li et al. 2024). A key step in advancing prompt multiplicity research is extending beyond structured outputs to encompass the open-ended nature of GenAI usage.

References

Aerni, M.; Rando, J.; Debenedetti, E.; Carlini, N.; Ippolito, D.; and Tramèr, F. 2025. Measuring Non-Adversarial Reproduction of Training Data in Large Language Models. In

The Thirteenth International Conference on Learning Representations.

Amatriain, X. 2024. Prompt design and engineering: Introduction and advanced methods. *arXiv preprint arXiv:2401.14423*.

Black, E.; Koepke, L.; Kim, P.; Barocas, S.; and Hsu, M. 2024. The Legal Duty to Search for Less Discriminatory Algorithms. *arXiv preprint arXiv:2406.06817*.

Black, E.; Raghavan, M.; and Barocas, S. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *FAccT*.

Chataigner, C.; Ma, R.; Ganesh, P.; Taïk, A.; Creager, E.; and Farnadi, G. 2025. Say It Another Way: A Framework for User-Grounded Paraphrasing. *arXiv preprint arXiv:2505.03563*.

Creel, K.; and Hellman, D. 2022. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1): 26–43.

Dhar, V. 2024. The paradigm shifts in artificial intelligence. *Communications of the ACM*, 67(11): 50–59.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Ganesh, P. 2024. An Empirical Investigation into Benchmarking Model Multiplicity for Trustworthy Machine Learning: A Case Study on Image Classification. In *2024 IEEE/CVF WACV*. IEEE.

Ganesh, P.; Chang, H.; Strobel, M.; and Shokri, R. 2023. On The Impact of Machine Learning Randomness on Group Fairness. In *FAccT*.

Ganesh, P.; Shokri, R.; and Farnadi, G. 2025. Rethinking Hallucinations: Correctness, Consistency, and Prompt Multiplicity. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Gomez, J. F.; Machado, C. V.; Paes, L. M.; and Calmon, F. P. 2024. Algorithmic Arbitrariness in Content Moderation. *arXiv preprint arXiv:2402.16979*.

Hsu, H.; and Calmon, F. 2022. Rashomon capacity: A metric for predictive multiplicity in classification. *Advances in Neural Information Processing Systems*, 35: 28988–29000.

Hsu, H.; Li, G.; Hu, S.; et al. 2024. Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Estimation. *arXiv preprint arXiv:2402.00728*.

Jain, S.; Creel, K.; and Wilson, A. C. ???? Position: Scarce Resource Allocations That Rely On Machine Learning Should Be Randomized. In *Forty-first International Conference on Machine Learning*.

Kulynych, B.; Hsu, H.; Troncoso, C.; and Calmon, F. P. 2023. Arbitrary decisions are a hidden cost of differentially private training. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1609–1623.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; et al. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Long, C.; Hsu, H.; Alghamdi, W.; and Calmon, F. 2023. Individual Arbitrariness and Group Fairness. *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; and Stenetorp, P. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8086–8098.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive Multiplicity in Classification. *Proceedings of the 37th International Conference on Machine Learning, PMLR*.

More, Y.; Ganesh, P.; and Farnadi, G. 2024. Towards more realistic extraction attacks: An adversarial perspective. *arXiv preprint arXiv:2407.02596*.

Nasr, M.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A. F.; Ippolito, D.; Choquette-Choo, C. A.; Wallace, E.; Tramèr, F.; and Lee, K. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Sclar, M.; Choi, Y.; Tsvetkov, Y.; and Suhr, A. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Shi, F.; Chen, X.; Misra, K.; Scales, N.; Dohan, D.; Chi, E. H.; Schärli, N.; and Zhou, D. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 31210–31227. PMLR.

Song, Y.; Liu, X.; and Shou, M. Z. 2024. Diffsim: Taming diffusion models for evaluating visual similarity. *arXiv preprint arXiv:2412.14580*.

Voronov, A.; Wolf, L.; and Ryabinin, M. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. *arXiv preprint arXiv:2401.06766*.

Watson-Daniels, J.; Barocas, S.; Hofman, J. M.; and Chouldechova, A. 2023. Multi-Target Multiplicity: Flexibility and Fairness in Target Specification under Resource Constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 297–311.

Watson-Daniels, J.; Parkes, D. C.; and Ustun, B. 2023. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10306–10314.

Wu, F.; Black, E.; and Chandrasekaran, V. 2024. Generative monoculture in large language models. *arXiv preprint arXiv:2407.02209*.

Yuan, Y. 2023. On the power of foundation models. In *ICML*, 40519–40530. PMLR.