Scaling Mutual Predictability Training Beyond Static Datasets

Vinesh Nathan[1], Anish Basety[1], Ishraq Quayyum[1], Zachary Guzman[1]

Abstract

The Unsupervised Evaluation (UE) framework demonstrates that mutual predictability between model outputs serves as a strong proxy for truthfulness, matching supervised labels while capturing both consistency and informativeness. However, UE's adversarial training component is limited to selecting questions from finite static datasets. We propose a fundamental extension: replacing finite question selection with a co-trained adversarial language model that generates collections of interrelated questions from scratch. This generative approach addresses the core limitation of static dataset dependence while potentially achieving superior truth-seeking behavior. Our framework trains an adversary to produce maximally informative question collections that challenge the main model's consistency, creating a dynamic training environment where both models evolve together. We address the critical concern of whether consistency-focused training might harm exploration by demonstrating that informativeness requirements naturally encourage diverse reasoning patterns. Building on preliminary implementation results showing correlation with Arena ratings, we propose comprehensive experiments across TruthfulQA, GSM8K, and HotpotQA to validate that generative adversarial consistency training outperforms the original finite-selection UE approach.

1 Introduction

The Unsupervised Evaluation (UE) framework by Wen et al. [2024] represents a breakthrough in alignment research by demonstrating that mutual predictability between model outputs correlates strongly with human preference and factual accuracy. The key insight revolves around evaluating model A's outputs on collections of related questions by testing whether another model B, when conditioned on answers to some questions in the collection, assigns high log-probability to A's outputs on the remaining questions.

This mutual predictability metric captures not merely consistency but also informativeness. An uninformative response, while potentially correct, fails to provide sufficient context for predicting answers to related or downstream questions. This dual requirement creates a natural filter for high-quality, truthful responses that maintain logical coherence across question collections.

The UE paper includes an adversarial training component where an adversary selects question subsets from finite datasets to maximize the mutual predictability signal. However, this approach faces fundamental limitations due to its dependence on pre-existing question pools. The adversary's effectiveness is constrained by the diversity and quality of available questions, potentially limiting the richness of the training signal and the model's ability to handle novel reasoning patterns.

Our research addresses this limitation through a fundamental extension: we replace the finite question selection mechanism with a co-trained adversarial language model that generates collections of interrelated questions from scratch. This generative approach eliminates dataset size constraints while enabling the adversary to continuously adapt its question generation strategy as the main model improves. Since the original UE approach demonstrated remarkable effectiveness, we hypothesize that this generalized version will achieve even better truth-seeking performance by providing richer and more diverse training signals.

2 Core Methodology: From Selection to Generation

2.1 Understanding the UE Baseline

The original UE framework establishes mutual predictability as a training signal through the following procedure. Given a dataset of question-answer pairs, an adversary selects subsets that maximize the mutual predictability score when answered by the main model. The mutual predictability for a collection $Q = \{(q_i, a_i)\}_{i=1}^{k}$ is computed as:

$$\mathrm{MP}(Q) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j \neq i} \log p(a_j | q_j, Q_{-j}) \tag{1}$$

where Q_{-j} represents all question-answer pairs except the j-th one. This evaluation can be extended to cross-model scenarios using a matrix formulation where rows represent models making predictions and columns represent models providing context, enabling systematic analysis of inter-model consistency patterns. The adversary's role involves discrete optimization over finite question pools to identify subsets that produce high mutual predictability scores, effectively creating challenging yet coherent evaluation scenarios for the main model.

While this approach demonstrates impressive results, the discrete selection process faces scalability limitations. The adversary cannot create entirely new questions or explore reasoning patterns beyond those present in the original dataset. This constraint becomes particularly problematic when dealing with specialized domains or emerging knowledge areas where existing question datasets may be incomplete or outdated.

2.2 Generative Adversarial Extension

Our proposed extension replaces the discrete selection adversary with a generative language model G_θ that produces entire question collections from scratch. The generative adversary can be optimized to either increase or decrease mutual predictability depending on the training objective, with current research exploring which direction yields superior alignment results. The primary objective formulation maximizes mutual predictability:

$$\mathcal{L}_G = -\mathbb{E}_{Q_g \sim G_\theta}[\mathrm{MP}(Q_g, \mathrm{LM}_\phi(Q_g))] \tag{2}$$

where Q_g represents a generated question collection, and $\mathrm{LM}_\phi(Q_g)$ denotes the main model's answers to these questions. However, alternative formulations that minimize agreement may also provide valuable training signals by forcing models to handle challenging edge cases and contradictory scenarios. The adversary learns to generate question sequences that maximize the mutual predictability signal while maintaining semantic coherence and factual grounding.

This generative approach offers several advantages over finite selection. First, the adversary can create questions tailored to specific weaknesses or knowledge gaps in the main model, leading to more targeted training signals. Second, the infinite space of possible questions enables continuous adaptation as both models evolve during training. Third, the adversary can explore novel question formulations and reasoning patterns that may not exist in static datasets.

The training procedure involves alternating optimization between the main model and the generative adversary. The adversary generates question collections designed to maximize mutual predictability, while the main model learns from high-scoring question-answer pairs through supervised fine-tuning or reinforcement learning approaches.

2.3 Addressing the Exploration Concern

A critical concern regarding consistency-based training involves its potential impact on model exploration capabilities. Since consistency requirements appear strongly exploitation-oriented, they might discourage the model from exploring diverse reasoning patterns or considering alternative perspectives. This concern is particularly relevant for generative models where creativity and diverse output generation are often desired properties.

However, the informativeness component of mutual predictability naturally addresses this concern. Questions that elicit uninformative responses receive low mutual predictability scores because such responses fail to provide useful context for predicting answers to related questions. This creates an implicit pressure toward informative, substantive responses that demonstrate genuine understanding rather than safe but vacuous outputs.

Furthermore, the generative adversary's ability to create diverse question collections ensures that the main model encounters a wide range of reasoning challenges. Rather than converging to a narrow set of response patterns, the model must maintain flexibility to handle the continuously evolving question distributions produced by the adversary. This dynamic training environment naturally encourages robust reasoning capabilities that generalize across diverse domains and question types.

3 Experimental Framework and Validation

3.1 Building on Preliminary Results

Our research builds upon preliminary implementation work that has already demonstrated promising results. The team has successfully implemented the core UE pipeline using approximately 300 question-answer pairs from TruthfulQA, GSM8K, and HotpotQA datasets, with round-robin log-probability calculations implemented in Google Colab. The implementation includes selection of the most consistent 100 questions and cross-model mutual predictability evaluation showing correlation with Arena ratings. These results provide strong empirical support for the mutual predictability framework and establish a solid foundation for extending to generative adversarial training.

The preliminary work also includes regression analysis confirming that the mutual predictability correlation with Arena ratings cannot be explained by common confounders such as response length or stylistic factors. This finding strengthens the case for mutual predictability as a genuine measure of response quality and truthfulness rather than a proxy for superficial characteristics. Current implementation bottlenecks center on compute resource availability rather than conceptual or coding challenges, with the team awaiting RunPod access for large-scale fine-tuning experiments using LLaMA models with LoRA or QLoRA methods.

3.2 Comparative Experimental Design

Our experimental validation centers on direct comparison between the original finite-selection UE approach and our generative adversarial extension. This comparison isolates the specific contribution of generative question creation while controlling for other methodological factors. We evaluate both approaches across multiple model architectures including LLaMA-2-7B, Mistral-7B, and other frontier models to ensure robust generalization.

The evaluation protocol encompasses multiple complementary metrics to provide comprehensive assessment of model consistency and truthfulness. Primary metrics include mutual predictability scores computed through round-robin evaluation, agreement levels measuring how frequently models produce similar or identical answers, and KL divergence analysis to quantify the distance

between probability distributions of different model outputs. We measure improvements in TruthfulQA accuracy, GSM8K mathematical reasoning, and HotpotQA multi-hop reasoning to assess whether generative adversarial training produces genuine improvements in reasoning capabilities. Additionally, we conduct transfer studies to examine whether models trained using our approach maintain their advantages when evaluated on entirely novel question distributions.

Cross-model evaluation provides another validation dimension by testing whether mutual predictability improvements achieved through generative adversarial training transfer across different model architectures. This analysis helps distinguish between approach-specific gains and fundamental improvements in reasoning consistency and truthfulness.

3.3 Informativeness and Exploration Analysis

To directly address concerns about potential harm to exploration capabilities, we design specific experiments that measure both consistency and diversity in model outputs. We evaluate the models' ability to generate diverse yet coherent responses to open-ended questions, assess their performance on creative reasoning tasks, and analyze the distribution of response types across different question categories.

The informativeness analysis examines whether responses generated by models trained using our approach provide sufficient context for answering related downstream questions. This evaluation directly tests the core hypothesis that mutual predictability captures informativeness as well as consistency. We measure downstream question answering accuracy when models are provided with context from previous responses, comparing performance between baseline models and those trained using generative adversarial consistency training.

We also conduct detailed analysis of the question collections generated by the adversarial model to understand the types of reasoning patterns and knowledge domains that emerge during training. This analysis provides insight into whether the generative adversary successfully identifies diverse challenging scenarios or converges to narrow question types that might limit the main model's learning.

4 Implementation Strategy and Resource Requirements

4.1 Technical Architecture

The implementation leverages the team's existing Google Colab Pro infrastructure supplemented by RunPod A100 GPU instances for computationally intensive training phases. The generative adversary development will incorporate research into existing question-generating and puzzle-generating fine-tuned language models, particularly open-source LLaMA-based architectures that can be adapted for adversarial question collection generation.

Parameter-efficient fine-tuning through Low-Rank Adaptation (LoRA) and Quantized LoRA (QLoRA) enables practical training within computational constraints while maintaining effectiveness. The alternating optimization procedure requires careful scheduling to ensure stable convergence of both the main model and generative adversary. We implement gradient accumulation and mixed-precision training to maximize computational efficiency within available resource limitations.

The question generation process incorporates diversity constraints and semantic coherence checks to ensure that generated collections maintain high quality while exploring diverse reasoning patterns. Human review protocols will be implemented to verify generated questions for correctness and coherence, with manual verification serving as the primary quality control mechanism for catching hallucinations and factual errors. This human review component will be clearly documented in

the methodology section to ensure reproducibility and transparency.

4.2 Evaluation and Validation Pipeline

The evaluation pipeline extends the existing implementation to support continuous assessment of both models throughout the co-training process. We track mutual predictability scores, downstream task performance, and diversity metrics to ensure that training progress aligns with our theoretical predictions. Regular checkpointing enables detailed analysis of learning dynamics and identification of potential training instabilities.

Automated evaluation against held-out test sets provides objective performance measurement, while human evaluation studies assess subjective response quality and informativeness. The combination of quantitative metrics and qualitative assessment ensures comprehensive validation of the approach's effectiveness across multiple evaluation dimensions.

Integration with existing benchmark datasets enables direct comparison with baseline approaches while maintaining compatibility with established evaluation protocols. This design choice facilitates reproduction of results and enables fair comparison with alternative alignment methodologies.

5 Expected Outcomes and Impact

5.1 Performance Predictions

Based on the strong performance of the original UE framework and our preliminary results, we anticipate significant improvements over finite-selection adversarial training. Specifically, we expect the generative approach to achieve 5-10 point improvements in TruthfulQA accuracy and similar gains in mathematical reasoning benchmarks. These improvements should result from the richer training signal provided by dynamically generated question collections that better target model weaknesses.

The correlation with Arena ratings should strengthen compared to baseline approaches, potentially achieving correlation coefficients exceeding those observed in the original UE work. This improvement would demonstrate that generative adversarial training produces models that better align with human preference judgments across diverse evaluation scenarios.

Generalization to novel question distributions represents a critical validation metric that distinguishes genuine reasoning improvements from dataset-specific overfitting. We expect models trained using our approach to maintain performance advantages when evaluated on entirely new question types and domains, indicating robust improvements in underlying reasoning capabilities.

5.2 Theoretical Contributions

This research establishes the first systematic framework for scaling mutual predictability training beyond static dataset constraints. The theoretical analysis of generative adversarial consistency training provides new insights into the relationship between consistency, informativeness, and truthfulness in language model behavior. Our work demonstrates that consistency requirements, when properly formulated, enhance rather than constrain model capabilities.

The exploration versus exploitation analysis addresses fundamental questions about the impact of consistency training on model diversity and creativity. By demonstrating that informativeness requirements naturally encourage diverse reasoning patterns, we provide theoretical and empir-

ical evidence that consistency-based training can improve model capabilities without sacrificing flexibility.

5.3 Practical Applications

The successful demonstration of generative adversarial consistency training would provide a scalable alternative to human preference collection for alignment applications. Organizations developing language models could adopt this framework to improve truthfulness and consistency without the substantial costs associated with human feedback collection. The approach's dataset-agnostic nature enables application across diverse domains and languages without requiring manual dataset curation.

The toolkit resulting from this research enables continuous adaptation to emerging knowledge domains and question types. As new information becomes available or novel reasoning patterns emerge, the generative adversary can automatically incorporate these developments into the training process without requiring manual intervention or dataset updates.

6 Conclusion

This research proposes a fundamental extension to the successful UE framework by replacing finite question selection with generative adversarial question creation. The approach addresses core limitations of static dataset dependence while leveraging the proven effectiveness of mutual predictability as a training signal. Our preliminary results provide strong empirical foundation for this extension, and the proposed experiments will validate whether generative adversarial consistency training achieves superior truth-seeking performance compared to finite-selection approaches.

The theoretical analysis of consistency versus exploration concerns demonstrates that informativeness requirements naturally encourage diverse reasoning patterns, addressing potential criticisms of consistency-focused training. The practical implications of this work extend beyond academic research to provide scalable alignment methodologies for real-world language model deployment.

By building directly on the established success of the UE framework while addressing its fundamental limitations, this research represents a natural evolution in mutual predictability-based training that maintains theoretical grounding while expanding practical applicability. The proposed experiments will provide definitive evidence for the effectiveness of generative adversarial consistency training and establish new directions for unsupervised alignment research.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Wen, L., Shi, W., Jang, J., Dong, Y., Li, K., Wu, J., Xie, T., Liu, P., Kim, H., Zhao, D., et al. Unsupervised evaluation of language models. *arXiv preprint arXiv:2506.10139*, 2024.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.