

Idempotent GPTs act may provide robustness by design^{*}

Sergei Kudriashov Jamilya Erkenova
HSE *Independent Researcher*

With
PIBBSS, Timaeus, & Apart Research

Abstract

Idempotence is one of the central concepts in quantum physics, corresponding to an operator that doesn't change its output being applied twice. Enforcing idempotence in generative deep learning may be interpreted as imposing a constraint on the model to be a projector on the manifold, corresponding to the train-time target distribution, which was explored for image generation models by Shocher et al. 2023. Idempotent test-time training has predicted to be a valuable approach for uncertainty quantification and adaptation to distribution shifts Durasov et al. 2025. We find that although language models are iterative refiners of token predictions, they struggle to preserve idempotence. Thus, we train small-scale idempotent GPT model with expected qualities by design and provide proof-of-concept code for evaluations. Our code is available at [\[github.com/erkenovaj/ittlpe\]\(https://github.com/erkenovaj/ittlpe\)](https://github.com/erkenovaj/ittlpe)

Keywords: Physics-informed AI safety, [your specific physics approach], [AI safety problem area], [your methodology]

1 Introduction

When it comes to neural networks, especially Large Language Models(LLM), any unintended behavior or uncertainty can lead to catastrophic consequences. Adversarial robustness is a crucial quality that a model should possess. Having the ability to accurately estimate the probability of dangerous behavior and manage the model to lower its uncertainty, especially in out-of-distribution or distribution shift scenarios, is an important research area. Default adversarial robustness methods, such as adversarial training, are not effective for such scenarios. The reason is that if it is difficult to find an input that can produce the harmful output, then the methods are likely to fail (Webb et al. 2019).

A possible approach to decreasing possible risk and improving model's adaptation to distribution shifts is test-time training. Test-time training performs an auxiliary self-supervised task on each test sample during inference (Durasov et al. 2025). This way, the model is enabled to handle out-of-distribution data using only the test batch. Recently researches linking adversarial robustness and perceptually aligned gradients to manifold hypothesis emerged. Essentially, if the network maps

^{*}Research conducted at the AI Safety x Physics Grand Challenge, 2025

the input onto the manifold, to lower the probability of the unwanted behavior, we would want the data to keep being on the same manifold. In physics and math, it corresponds to the idea of idempotence. The concept of idempotence—where an operation yields the same result no matter how many times it is applied—finds interesting applications in quantum physics, particularly in the formalism of quantum mechanics and quantum information theory.

Idempotence can also be used to ensure that the models behave as intended. This strategy results in a model capable of generating an output in one step, maintaining a consistent latent space, while also allowing sequential applications for refinement.

In the context of deep learning idempotency requires the following conditions (Shocher et al. 2023): 1. Instances from the target distribution should map to themselves, namely $f(x) = x$. We define the target manifold as the set of all instances that f maps to themselves. 2. Instances that form the source distribution should map onto the defined target manifold. This is achieved by optimizing the idempotence term, $f(f(z)) = f(z)$ which encourages the range of $f(z)$ to be on the target manifold. Under ideal assumptions such a process provably converges to the target distribution.

Although those ideas were linked in the paper Idempotent Test-Time Training (ref), they only were tested on CNNs and on algebraic scenarios. Knowing that transformer models, such as BERT or GPT, are built in a way to iteratively refine their prediction (Devlin et al. 2019; Brown et al. 2020), we think that it is possible to reinforce idempotency in such models.

The problem area this research is associated with is Scaling Laws, Phase Transitions, and Emergence. Concepts such as rare behaviors and idempotency can be observed in different physical systems. Using phenomena seen in physics, we hope to bridge the gap for low probability estimations. How does it advance our theoretical understanding or practical capabilities in ensuring safe AI systems? Dealing with rare behaviors prevents models from showing harmful patterns.

Briefly explain the physics concepts, methodologies, or frameworks that form the foundation of your approach.

2 Methods

Idempotence is a quality of a physical concept, it is rarely reinforced in physics, but can be observed. For example, quantum decoherence and zeno effect, which states that the quantum Zeno effect (where frequent measurements freeze evolution) relies on repeated projective (idempotent) measurements. This way qubits are stabilized against decoherence. Some quantum channels (CPTP maps) have idempotent components, leading to noiseless subsystems or decoherence-free subspaces. Some quantum channels (CPTP maps) have idempotent components, leading to noiseless subsystems or decoherence-free subspaces. So, as part of the research we tried to observe idempotency in some LM models and also building a toy-model that is specifically idempotent.

The first experimental setup:

- 100 samples from the TriviaQA validation set, used as prompts
- Models: SmolLM3-3B, Qwen2-7B, Llama-2-7b, Mistral-7B-v0.1

First, we predict the next token. Then, we additionally take the hidden state for the predicted model and run the model again on it, testing if output is the same. Iteratively, we add more and more tokens. After n tokens (we chose 100) we calculate KL-divergence to see if the distributions are similar.

The second experimental setup:

We train a small GPT model based on nanoGPT Karpathy 2022 on "shakespear-char" dataset with the standard next token prediction objective, with an additional objective of making the model idempotent on out-of-distribution samples.

Taking inspiration from Shocher et al. 2023, we utilize the same approach to model training. However, we utilize a random sample of tokens as out-of-distribution sample to make it idempotent. We also exchange L2 distance loss between predicted values for KL-divergence between distributions as we are working with probability distributions over discrete domains.

The model appears to be capable of generating fluent text from the domain and attains low loss values, which makes its further development a promising research direction.

3 Results

This research was motivated by low probability estimation. The key findings are:

- **Theoretical Insights:** A new framework that combines idemotency and test-time training in LLM in order for the model to be more robust against out-of-distribution tokens.
- **Empirical Findings:** We checked idempotency of several LLMs. The results against a random baseline are shown in the 1. We can see, that most models such as are unable to show better results than the random baselines, while models actually show something resembling idempotency. This finding can be used in the future to explicitly est if the idempotent models predict rare behaviors more accurately. It also shows that LLM models are not idempotent by themselves.

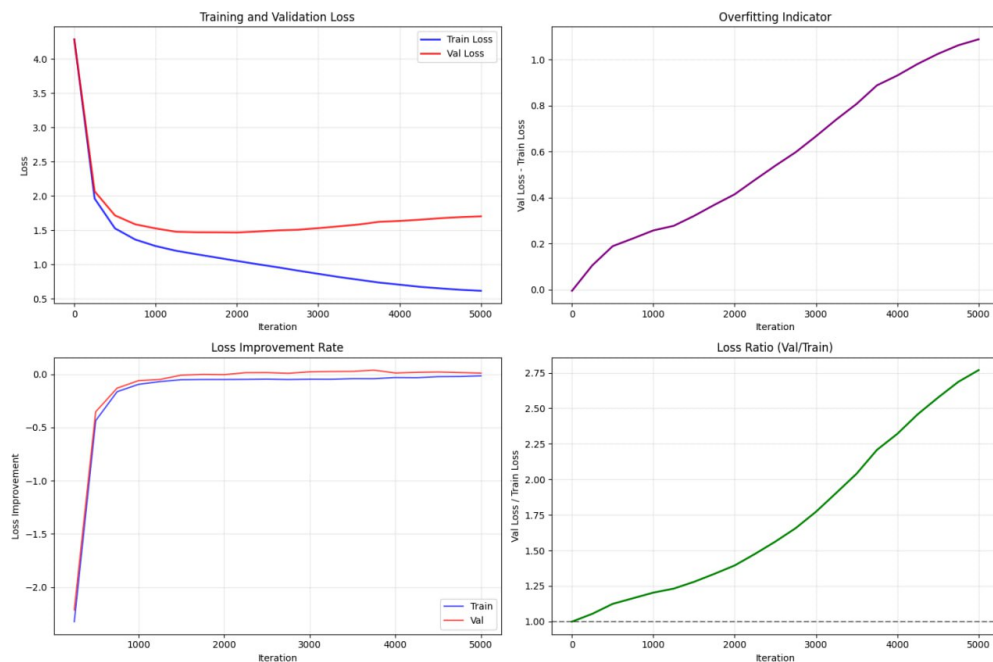


Figure 1: Training metrics of the GPT model

We also trained a GPT-like model to have idempotency. Unfortunately, we did not have enough time to test it the same way we did with other models. But it also provides a good base for future work.

- **Methodological Contributions:** New approaches for applying physics concepts to AI safety, ability to adapt to distribution shifts at test-time and possibly lower catastrophic risks related to deployment scale inference.
- **Proof-of-Concept:** GPTs are capable of being idempotent projectors on the target manifold. The training process is shown on 2



Figure 2: Training metrics of the GPT model

4 Discussion and Conclusion

This study was motivated by the challenge of accurately estimating low-probability events in language models.

These insights lay the groundwork for future research in improving model stability, safety, and generalization under distributional uncertainty.

5 References

References

- Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL]. URL: <https://arxiv.org/abs/2005.14165>.
- Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL]. URL: <https://arxiv.org/abs/1810.04805>.
- Durasov, Nikita et al. (2025). *IT³: Idempotent Test-Time Training*. arXiv: 2410.04201 [cs.CV]. URL: <https://arxiv.org/abs/2410.04201>.
- Jones, Erik et al. (2025). *Forecasting Rare Language Model Behaviors*. arXiv: 2502.16797 [cs.LG]. URL: <https://arxiv.org/abs/2502.16797>.

Karpathy, Andrej (2022). *NanoGPT*. <https://github.com/karpathy/nanoGPT>.
 Shocher, Assaf et al. (2023). *Idempotent Generative Network*. arXiv: 2311.01462 [cs.CV]. URL: <https://arxiv.org/abs/2311.01462>.
 Webb, Stefan et al. (2019). *A Statistical Approach to Assessing Neural Network Robustness*. arXiv: 1811.07209 [stat.ML]. URL: <https://arxiv.org/abs/1811.07209>.
 Wu, Gabriel and Jacob Hilton (2025). *Estimating the Probabilities of Rare Outputs in Language Models*. arXiv: 2410.13211 [cs.LG]. URL: <https://arxiv.org/abs/2410.13211>.

6 Appendix

Additional Materials: Include any supplementary theoretical derivations, extended results, code snippets, or detailed mathematical frameworks that support your main findings.

If you used large language models (ChatGPT, Claude, etc.) during your research process, please declare their usage here. Include what tasks they assisted with (e.g., literature review, code debugging, mathematical derivations, writing assistance) and any specific prompts or approaches that were particularly helpful for your physics-AI safety research.