inteli instituto
de tecnologia
e liderança

INSTITUTE OF TECHNOLOGY AND LEADERSHIP

COMPUTER SCIENCE

MELYSSA DE SOUZA GONZALES ROJAS

Training and Replication of Collapsed Models

São Paulo

2025

MELYSSA DE SOUZA GONZALES ROJAS

Training and Replication of Collapsed Models

Course conclusion paper submitted to the Institute of Technology and Leadership (Inteli), as a partial requirement for obtaining the degree of Bachelor in Computer Science.

Advisor: Prof. Dr. Rafael Will Macedo de Araujo

São Paulo
2025

ABSTRACT

This research explores the phenomenon of language model collapse, where large models trained on synthetic data generated by their predecessors experience a progressive decline in response quality. As data distributions become increasingly concentrated on high-probability tokens, less frequent words and contexts are underrepresented, impairing the models' ability to generate coherent responses to diverse prompts. The study focuses on replicating prior experiments to construct collapsed models, providing a foundation for future analysis involving autonomous agents designed to mitigate this degradation. Evaluation included perplexity analysis and qualitative assessment of responses across different model generations, revealing increased uncertainty, repetition, and hallucinations in later generations. The results confirm the expected collapse pattern and highlight the risks of over-reliance on synthetic data. As a next step, the research proposes the use of autonomous agents focused on guardrails and Retrieval-Augmented Generation (RAG) techniques to control and reduce degenerated responses in user dialogue contexts.

Keywords: language model collapse, synthetic data, perplexity, autonomous agents, guardrails, RAG, response degeneration, LLM evaluation

SUMMARY

1 INTRODUCTION

The phenomenon of collapse in Large Language Models (LLMs) refers to the process by which models trained successively on synthetic data generated by their predecessors begin to produce increasingly degenerated responses, regardless of the user's prompt context. With each generation, the training data becomes more concentrated on high-probability words, while less frequent words and contexts are gradually excluded from the datasets. This reduction in linguistic diversity hinders the model's ability to generate coherent and contextually appropriate responses, especially in uncommon or complex situations [1].

Given this concerning scenario, it becomes essential to develop control and mitigation mechanisms capable of addressing response degeneration in real user interactions. This research aims to advance in this direction through the development of autonomous agents focused on implementing guardrails, which are instructional components designed to limit inadequate outputs from the models [2]. The goal is to reduce the negative impact of collapse in practical applications.

At this stage of the study, the focus is on building collapsed models as an experimental foundation for future analysis with autonomous agents. These models were constructed by replicating experiments from a previous study, providing a technical basis for subsequent evaluations. The objective of this paper is to describe the main components used in building these models, to present a perplexity-based analysis that quantifies the difficulty the models face in generating tokens across different contexts, and to conduct a qualitative evaluation of responses based on selected prompts, observing the degree of collapse across model generations. Finally, the next steps of the research are outlined, aiming to apply and evaluate autonomous agents in mitigating collapse in LLMs.

2 MATERIALS AND METHODS

This section describes the materials and procedures used for the experimental replication of a previous study on the phenomenon of collapse in language models. The objective is to reproduce the conditions of the original research by detailing the model selection, dataset, fine-tuning process, synthetic dataset generation, and the infrastructure employed, ensuring the fidelity and reliability of the results obtained [1].

The model used was facebook/opt-125m, a causal language model pre-trained with 125 million parameters based on the Transformer architecture. This model was chosen based on the original study being replicated. The fine-tuning of the models was carried out in the Google Colab Pro+ environment using an A100 GPU, providing efficient training performance.

For fine-tuning, the WikiText-2 raw v1 dataset was tokenized using the tokenizer associated with the model and segmented into fixed blocks of 64 tokens to facilitate batch processing. The model was trained for five epochs using the AdamW optimizer with a learning rate of 2e-5. Training management was handled by PyTorch Lightning, including saving the best checkpoint based on validation loss to ensure optimal performance.

The generation of synthetic datasets for subsequent generations followed a recursive process. After training generation 0, the model was used to generate textual continuations for each sample in the original dataset, forming the dataset for generation 1. This procedure was repeated to produce datasets and train models for generations 2 and 3, maintaining a block size of 64 tokens and applying beam search with a beam width of five to improve the quality of the generated sequences.

The synthetic dataset generation process was conducted on robust hardware comprising an Intel Xeon Gold 64545 processor (2.2 GHz, 64 cores), 128 GB DDR5 RAM at 4800 MHz, and an Nvidia L4 GPU with 24 GB of memory, ensuring efficient and fast processing.

3 PERPLEXITY OF MODEL COLLAPSE

Perplexity is a widely adopted metric for evaluating the performance of language models, reflecting the model's level of uncertainty when predicting the next word in a text sequence. Intuitively, perplexity can be understood as the average number of words among which the model "hesitates" at each generation step. The lower the perplexity, the higher the model's confidence in its predictions:

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i \mid w_1, w_2, \ldots, w_{i-1})\right) \tag{1}$$

as defined in [3], metric is calculated based on the probabilities assigned to each word w_i in the sequence. For a sequence of N words, perplexity takes into account the model's "surprise" level regarding the occurrence of each word, given the previous context. For example, a perplexity of 5 indicates that, on average, the model considers five words to be equally probable at each step.

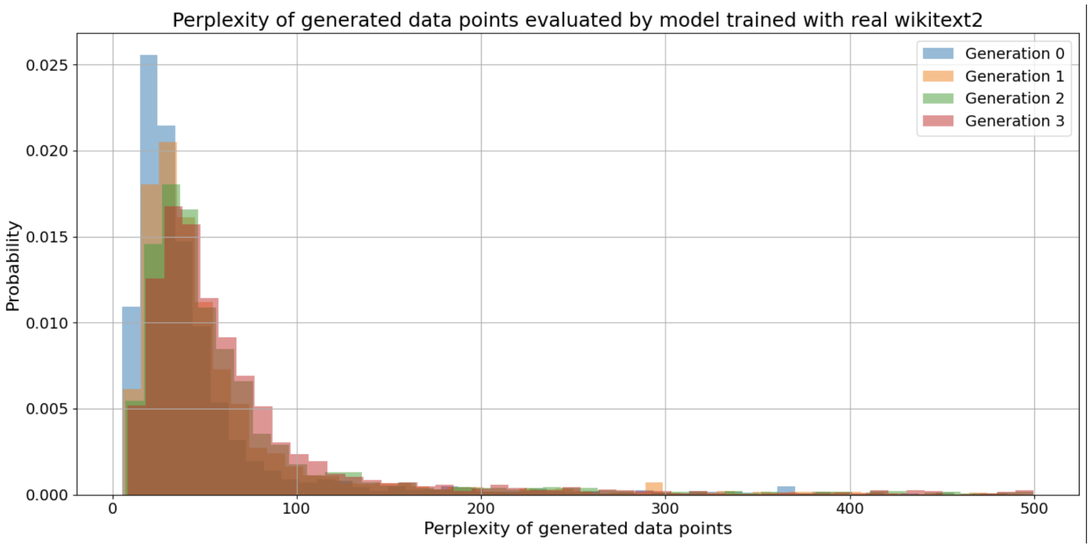

Figure 1: Perplexity distribution of texts generated across Generations 0 to 3, evaluated by a model trained on WikiText2. The X-axis represents perplexity scores (lower values indicate more natural text), and the Y-axis represents the probability density of texts within each perplexity range.

In Figure 1, shows the distribution of perplexity of texts generated by a language model across different generations, from Generation 0 to Generation 3. The evaluation was performed by a model previously trained on real data from the WikiText2 corpus, which serves as a reference to estimate how natural or plausible the generated texts are. Generation 0 exhibits the highest density of examples with low perplexity, indicating that these texts have a high similarity to the real data. This occurs because, at this stage, the texts are generated directly by the trained model without the accumulation of noise.

As the process advances to subsequent generations, specifically Generations 1, 2, and 3, the distribution progressively shifts toward higher perplexity values. This behavior indicates a degradation in text quality resulting from the model generating text based on its own previous outputs. This degradation pattern is clearly visible in the long tail of the distributions for later generations, suggesting that an increasing fraction of the texts have very high perplexity and are considered unlikely or unnatural by the evaluating model.

4 QUALITATIVE ANALYSIS OF MODEL RESPONSES TO PROMPTS

To demonstrate the quality of the models generated during the practical phase of this research, an evaluation of the responses was conducted based on seven selected prompts. A clear trend was observed across the model generations: generation 0 produced the most coherent responses, while starting from generation 1, signs of hallucinations appeared, and in subsequent generations, there was an increase in word repetition. This phenomenon of hallucination and repetition is associated with the high degree of uncertainty the models have when generating tokens from prompts, especially in less common use cases [1]. In such scenarios, the model tends to hallucinate or repeat connectors and frequent terms in the English language. It is important to highlight that as the data distribution loses diversity, there is a higher concentration of connectors, common everyday words, which reinforces the tendency toward increasingly less varied responses.

Model	Generated Response
Model 0 (Generation 0)	The secret to eternal life lies hidden within the soul.
Model 1 (Generation 1)	The secret to eternal life lies hidden within the soul of the person who dies. The soul of a person who dies is called the "" or or a a or the the or any any of of or of a the a any or of its its a any a its or of any other or anything else or in or to or you or
Model 2 (Generation 2)	The secret to eternal life lies hidden within the soul. The soul is the source of all all the the a a the all of of a and and all in the i i and its its all its a i all a its the and a all i a its i the its and of all and of and a in a
Model 3 (Generation 3)	The secret to eternal life lies hidden within the words words, or or any any or the or a a or in or to to or, a any of any other or anything else, any the any a the a to a, to any to, the to the to be or you or more or

Table 1: Responses generated by each model to the prompt: The secret to eternal life lies hidden within.

In Table 1, it is possible to see that the generation 0 model produces the most coherent response to the prompt, while models from subsequent generations show a growing repetition of common words, with the coherence of the responses decreasing as generations advance. These results confirm the degeneration of the model responses and indicate that, given the concentration of data in the fine-tuning dataset, increasingly focused on specific words and contexts, the model begins to prioritize more common terms, reducing its ability to produce diversified and contextually appropriate responses.

5 CONCLUSION

The practical phase of this research, focused on replicating experiments to construct collapsed language models, demonstrated that the generated models reached the expected level of degeneration, confirming the key aspects outlined in the foundational study. The perplexity analysis revealed increasing uncertainty in token prediction among later generations, impairing the coherence of the generated responses. Furthermore, the qualitative assessment showed a clear tendency toward repetition and the overuse of common words, even when disconnected from the intended meaning of the prompt, reinforcing the hypothesis of generational collapse.

These findings highlight that, without strict control over the balance between synthetic and human-authored data during training, future models are likely to exhibit worsening performance, with increasingly degenerated outputs [1]. In response, the next steps of this research involve the implementation and evaluation of autonomous agents, with a focus on the use of guardrails and Retrieval-Augmented Generation (RAG) techniques, to mitigate or block inadequate responses during user interactions. The ultimate goal is to assess and reduce the negative impact of collapsed models in real-world dialogue settings.

REFERENCES

[1] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[2] Salesforce. Einstein trust layer: Response journey. Technical Report Technical Report, Salesforce, 2025.

[3] Chao-Chung Wu, Zhi Rui Tam, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Clear minds think alike: What makes llm fine-tuning robust? a study of token perplexity. *arXiv preprint arXiv:2501.14315*, 2025.