NetSciX2026: International School and Conference on Network Science
February 17^{th}-20^{th}, 2026 - Auckland, New Zealand

Learning by Surprise: Surplexity for Mitigating Model Collapse in Generative AI

Keywords: LLM, Model Collapse, Linguistic Diversity, Synthetic Data, Artificial Intelligence

Extended Abstract

In recent years, generative AI has demonstrated remarkable advancements, particularly in conversational applications like ChatGPT, Google Gemini, Claude, and Llama [1]. A report by [2] predicts that by 2025, 90% of internet content will be AI-generated, increasing the chance that LGMs outputs may be used to train subsequent versions of these models. Recent research has highlighted the risks associated with a self-consuming loop, often referred to as *autophagy* process, where LGMs are recursively fine-tuned on their own outputs ([3]). Studies indicate that the process can lead to a phenomenon called *model collapse*, characterized by a significant loss of linguistic diversity. Subsequent research has explored several factors influencing model collapse, including model size, fine-tuning parameters, and data augmentation strategies. (see, e.g., [4–6]). However, our understanding of model collapse is a work in progress. As noted by [7], there are many alternative conceptions of collapse, and many open questions. In our work [8] we'll focus on three issues where progress is needed: (1) Model collapse is typically identified through metrics defined on *generated content*. But we do not yet have a way to characterize a collapsed model *in its own right*, as a probability model. (2) It is well known that fine-tuning on synthetic data can cause collapse—but at present we lack a principled understanding of which properties of training documents contribute to collapse. (3) At present, methods for mitigating collapse presuppose knowledge about whether training documents are synthetic or human-authored: an assumption that often does not hold in real-world settings.

We connect these three open questions. In response to (1), we propose a novel metric for a collapsed LLM, based on its own probability distributions that it generates, rather than the properties of the text it produces. We define a collapsed model as a model whose predicted probability distributions over the next word are *skewed*, measuring skewedness both with the Gini coefficient of the distribution and with an absolute threshold on its probability values. Using this new metric, we address issue (2), proposing a new definition of the class of documents responsible for collapse. By our definition, the key property of training documents that lead to collapse relates to the model's **surprise**: documents that are not surprising lead to collapse, while documents that are surprising do not. We formally define, given a model M and a document d the **surplexity** which measures how much the model is surprised about the document. We then use this new conception of surplexity to address issue (3): we propose a way to mitigate collapse by filtering training items by the level of surprise they elicit. In new experiments, we show our mitigation strategy is effective. crucially, the new strategy does not require knowledge about the synthetic or human-generated origin of training items. Interestingly, our new strategy is related to a conception of surprise-driven learning from cognitive science: this link may usefully connect the topic of model collapse with work on human learning mechanisms.

Furthermore, to better study the collapse phenomenon, we extend our characterization based on next-token probabilities into a network representation. Starting from a given prompt, we define as nodes the last token of the prompt and the tokens in the next-token predictions, connecting them with weighted edges where the weight corresponds to the token probability. We then randomly select one of the predicted tokens as the continuation of the prompt and

repeat this process for k steps. In this way, we obtain a network that captures not only the growing inequality in the probability distribution of a single prompt, but also how frequently the same token is suggested across different prompts as shown in Figure 1. This approach allows us to characterize model collapse in terms of network metrics and, additionally, provides a framework to compare the phenomenon of autophagy feedback loop in generative AI with other ecosystems [3].

References

[1] Junchao Wu et al. "A survey on llm-gernerated text detection: Necessity, methods, and future directions". In: *arXiv preprint arXiv:2310.14724* (2023).

[2] Europol Innovation Lab. *Facing Reality: Law Enforcement and the Challenge of Deepfakes*. 2021.

[3] Luca Pappalardo et al. "A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions". In: *arXiv preprint arXiv:2407.01630* (2024).

[4] Ilia Shumailov et al. "AI models collapse when trained on recursively generated data". In: *Nature* 631.8022 (2024), pp. 755–759.

[5] Sina Alemohammad et al. *Self-Consuming Generative Models Go MAD*. 2023.

[6] Yanzhu Guo et al. *The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text*. 2023.

[7] Rylan Schaeffer et al. *Position: Model Collapse Does Not Mean What You Think*. 2025. URL: https://arxiv.org/abs/2503.03150.

[8] Daniele Gambetta et al. *Learning by Surprise: Surplexity for Mitigating Model Collapse in Generative AI*. 2025. URL: https://arxiv.org/abs/2410.12341.

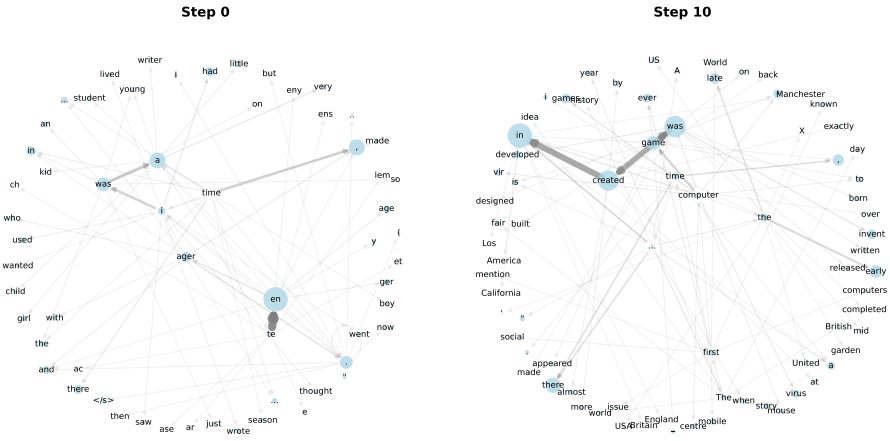

Figure 1: Network representation of the next-token probability distribution in the model at step 0 and step 10 of the autophagy process. Model collapse leads to an increased concentration of weight in the incoming links of a few nodes, reflecting the loss of linguistic diversity.