Model Collapse and the Need for Human-Generated Training Data

A Proposal for Certified Human-Written Training Data

Guillaume Lethuillier
glthr.com

August 5, 2025 · updated November 15, 2026

Abstract

I discuss the risk of training successive models on recursively generated data and propose verified human-written material produced by domain experts. The proposal is to use that material for fine-tuning and evaluation after initial pre-training, through either academic or commercial distribution. The certification process and the benefits of this approach remain proposals to test; this article does not demonstrate their effectiveness.

> **Update (November 2026):** The market for human-written training data that this article anticipates has grown considerably, with companies such as Scale AI, Surge AI, and Mercor recruiting domain experts, including PhDs, to produce training data for model creators.

Generative AI is poisoning its own well: online content is increasingly generated by AI; this data is used to train new models; those models then generate more online content, which in turn becomes training data. This creates a cycle that risks contaminating the sources AI relies on, potentially leading to diminished originality, amplified biases, and a disconnect from real-world information.

In 2024, researchers already raised a critical warning about this trend: "*The development of LLMs is very involved and requires large quantities of training data. Yet, although current LLMs . . . were trained on predominantly human-generated text, this may change. If the training data of most future models are also scraped from the web, then they will inevitably train on data produced by their predecessors.*"[1] The authors describe this as leading to a *model collapse*: "*Model collapse is a degenerative process affecting generations of learned generative models, in which the data they generate end up polluting the training set of the next generation. Being trained on polluted data, they then mis-perceive reality.*"[2]

I believe that we may soon need "certified" human-generated data to train models. In other words, model creators, to achieve the best model performances, may benefit from a service that provides training data that is guaranteed to come from human minds. This is because human-generated data possesses qualities currently difficult for AI to replicate: nuance, creativity, and common sense reasoning. Ensuring the source of training data will become increasingly important for them.

This idea raises multiple questions.

Cite as: Guillaume Lethuillier, "Model Collapse and the Need for Human-Generated Training Data", glthr.com, August 5, 2025. https://glthr.com/model-collapse-and-the-need-for-human-generated-training-data

[1] Shumailov, I., Shumaylov, Z., Zhao, Y. et al., AI models collapse when trained on recursively generated data. Nature 631, 755–759 (2024), p. 755.

[2] Ibid.

How Can We Ensure That Data Is Generated by Humans?

I see one robust approach: a "human-in-a-room" experiment. To guarantee human origin, participants should have no access to AI during data generation. The ideal setup would involve having people work in a controlled environment, such as a library. While establishing such an environment presents logistical challenges (cost, participant comfort), positive incentives (financial compensation or academic credit) could encourage participation.

Who Should Generate the Training Data?

Experts in their respective fields. PhDs, post-docs, doctors, mathematicians, historians, and others with deep subject matter knowledge. They would be asked to produce specific types of content (factual statements, creative writing prompts designed to test reasoning, or complex problem sets) with their knowledge, expertise, and available offline resources. To ensure consistent quality, a rigorous peer review process involving multiple experts would be essential. This project resembles the work of encyclopedists, *but instead of creating a first-order knowledge repository (an encyclopedia), they would create a higher-order one: high-quality training data for AI models.*

How Should This Data Be Made Available?

Two approaches are possible: open access, driven by academic contributions, or a commercial market where data is sold under licence and not publicly released. Given its potential value (significantly higher quality than readily available online datasets), I anticipate the emergence of a robust market for certified human-generated training data. However, a market-driven approach raises concerns about equitable access and could favor larger companies with greater financial resources.

Could AI Models Be Trained Purely on Human-Generated Data?

While humans cannot generate the sheer volume required for initial model training, a two-step approach is feasible. A thousand experts writing 1,500 words a day for 200 days would produce 300 million words: enough for fine-tuning, not for pre-training. Models could first be pre-trained on existing large datasets and then fine-tuned using certified human data. During this second stage, we could instruct the model to prioritize learning from the human dataset by weighting its loss function more heavily. A held-out part of this high-quality data would also serve as an invaluable test set for evaluating model performance and identifying biases.

How to cite this article

Guillaume Lethuillier, "Model Collapse and the Need for Human-Generated Training Data", glthr.com, August 5, 2025. https://glthr.com/model-collapse-and-the-need-for-human-generated-training-data

For BibTeX:

```
@misc{lethuillier2025model,
  author = {Lethuillier, Guillaume},
  title  = {Model Collapse and the Need for Human-Generated Training Data},
  year   = {2025},
  month  = aug,
  url    = {https://glthr.com/model-collapse-and-the-need-for-human-generated-training-data},
  note   = {glthr.com}
}
```

This document was typeset with LaTeX from the article published at https://glthr.com/model-collapse-and-the-need-for-human-generated-training-data, which remains the reference version (as updated on November 15, 2026).