SCUDEM Project Problem B
Drew Mastovsky, Peter Nolan, Mike Giles
Team 1129
Coach: Dr. Hasala Gallolu Kankanamalage
Roger WIlliams University, Bristol, RI USA

An AI Ouroboros Math Model

Brief Overview

Generative AI systems constantly learn from digital datasets. As this AI-content becomes more common online, these models risk training on their own AI outputs, leading to decline in accuracy known as *model collapse.*

Problem Statement: Model the dynamics of an ecosystem were multiple AI models interact with each other with new **human-generated content** and **AI generated content**

Preliminary Research

Core Question: How will the quality of content generated by an AI change when the proportion of human gen. vs AI gen. content within the training pool changes?

- We need to measure “quality”
- We need to see how it changes when the proportion changes
- Rate of positive injection (human content) and rate of elimination (model collapse)

Our System of Differential Equations

The Model Equations

$$\frac{dQ}{d\alpha} = -k\,Q(\alpha) + \lambda\,P_H(\alpha) \qquad \text{(Equation 1: Quality Dynamics)}$$

$$P_H(\alpha) = 1 - \alpha \qquad \text{(Equation 2: Proportion Constraint)}$$

Variable and Parameter Definitions

- $Q(\alpha)$: The **Content Diversity** (originality) of the content pool.
- $P_H(\alpha)$: The **Proportion of Human-Generated Content** in the pool.
- k: **Rate of Elimination** constant (strength of model collapse effect).
- λ: **Rate of Injection** constant (human content input).
- α: **Amount** of the AI content proportion.

Explicit Solution - Ready for fitting!

The Complete Solution

Substituting the constant C back into the general solution yields the complete, particular solution for the Content Diversity $Q(\alpha)$:

$$Q(\alpha) = \frac{\lambda}{k}\left(1 - \alpha + \frac{1}{k}\right) + \left(Q_0 - \frac{\lambda}{k}\left(1 + \frac{1}{k}\right)\right) e^{-k\alpha}$$

This equation defines the Content Diversity as a function of the AI content proportion, where Q_0 is the initial content diversity.

Generating Data - A Simulation Step-by-Step

1. Load real human data
   a. Uses the **AG News dataset**, a large human-written corpus of short news texts
   b. This serves as the *reference for high-quality, diverse text*
2. Generate Synthetic data
   a. Text is generated using **three different AI models** (TinyLlama, Phi-1.5, DistilGPT2) to represent different levels of AI output sophistication
   b. Each model produces hundreds of short “news headline” samples, mimicking how the internet might fill up with varying levels of synthetic content
3. Mix Human + AI Text
   a. Creates *mixtures* of human and AI texts controlled by a mixing parameter **α (alpha)**:
      i. α = 0 → 100% human
      ii. α = 1 → 100% AI
      iii. Intermediate α → mixed datasets
   b. Each mixture represents a different point in the “AI contamination” process

What Data is Gained?

- **Perplexity**: Measures predictability of text under a large “evaluation” language model
- (high perplexity -> chaotic language, low perplexity -> fluent, predictable text)

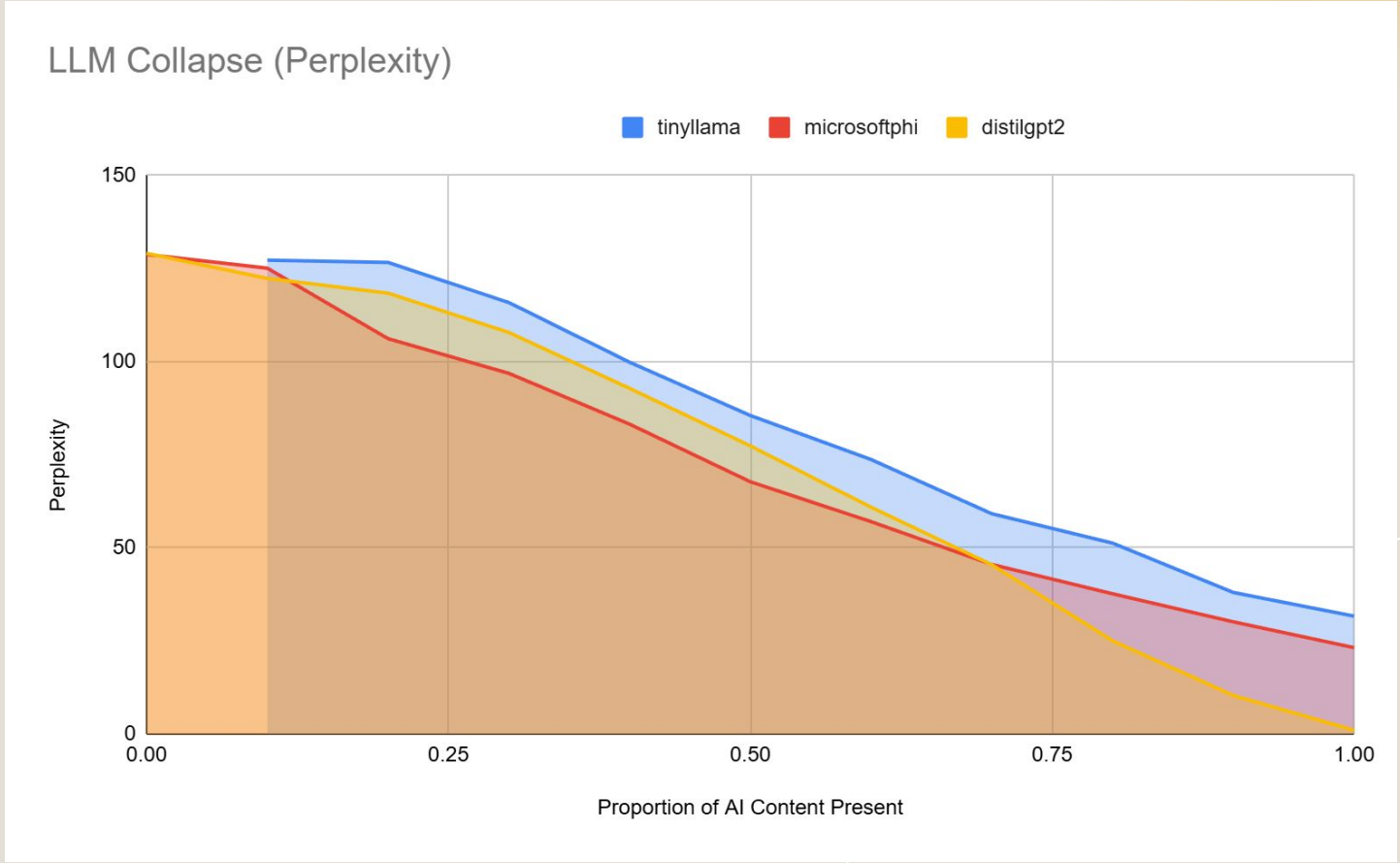
- **KL Divergence**: Quantifies how different the n-gram distributions of the mixed text are from human reference data
- (low kl -> similar to human text, high kl -> diverging linguistic patterns)

This will be used to measure LLM quality (Q())

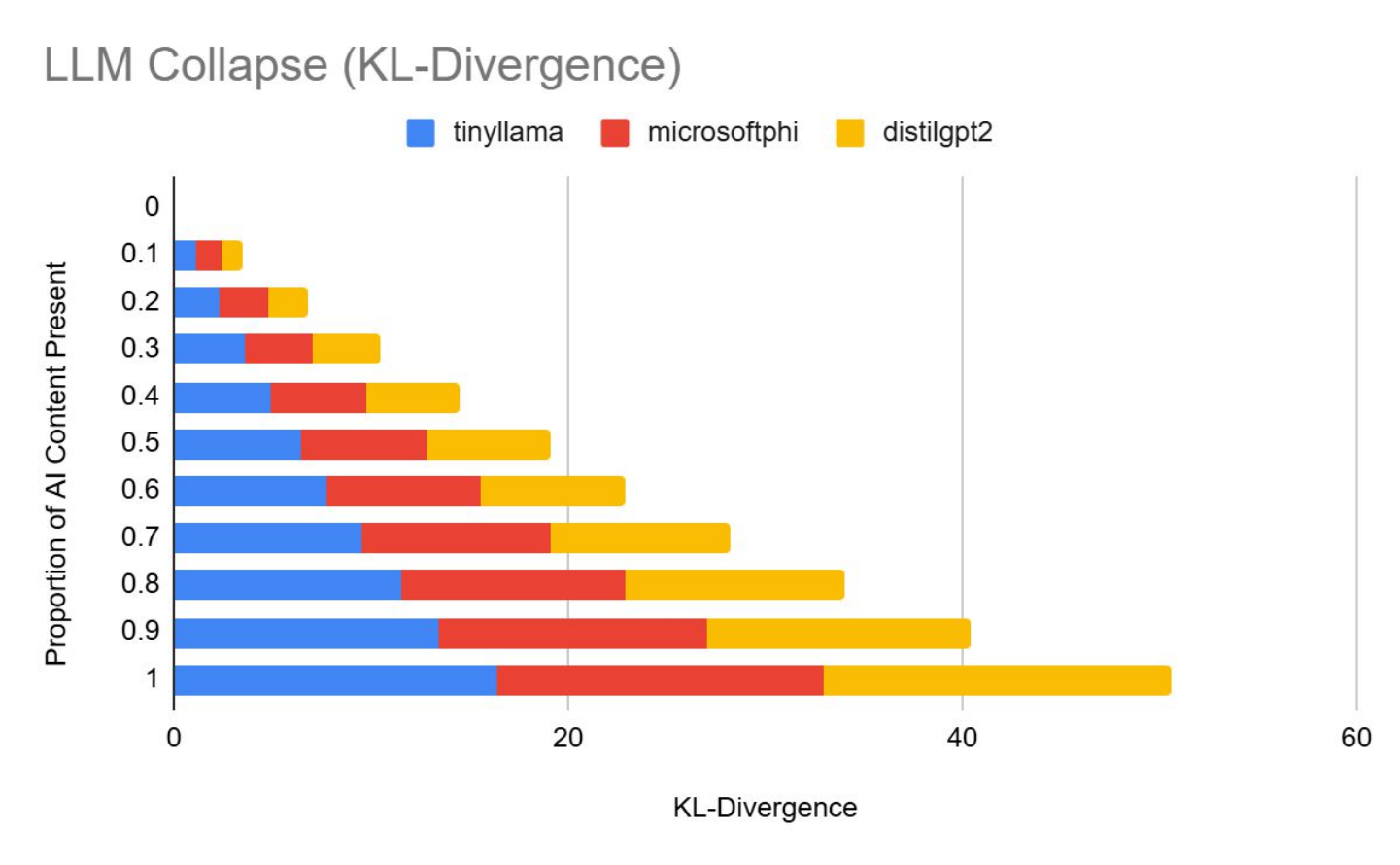
- **Distinct-2**: Measures lexical diversity
- (high d2 -> rich, varied vocabulary, low d2 -> repetitive, degenerative text)

**these were recorded for all three language models*

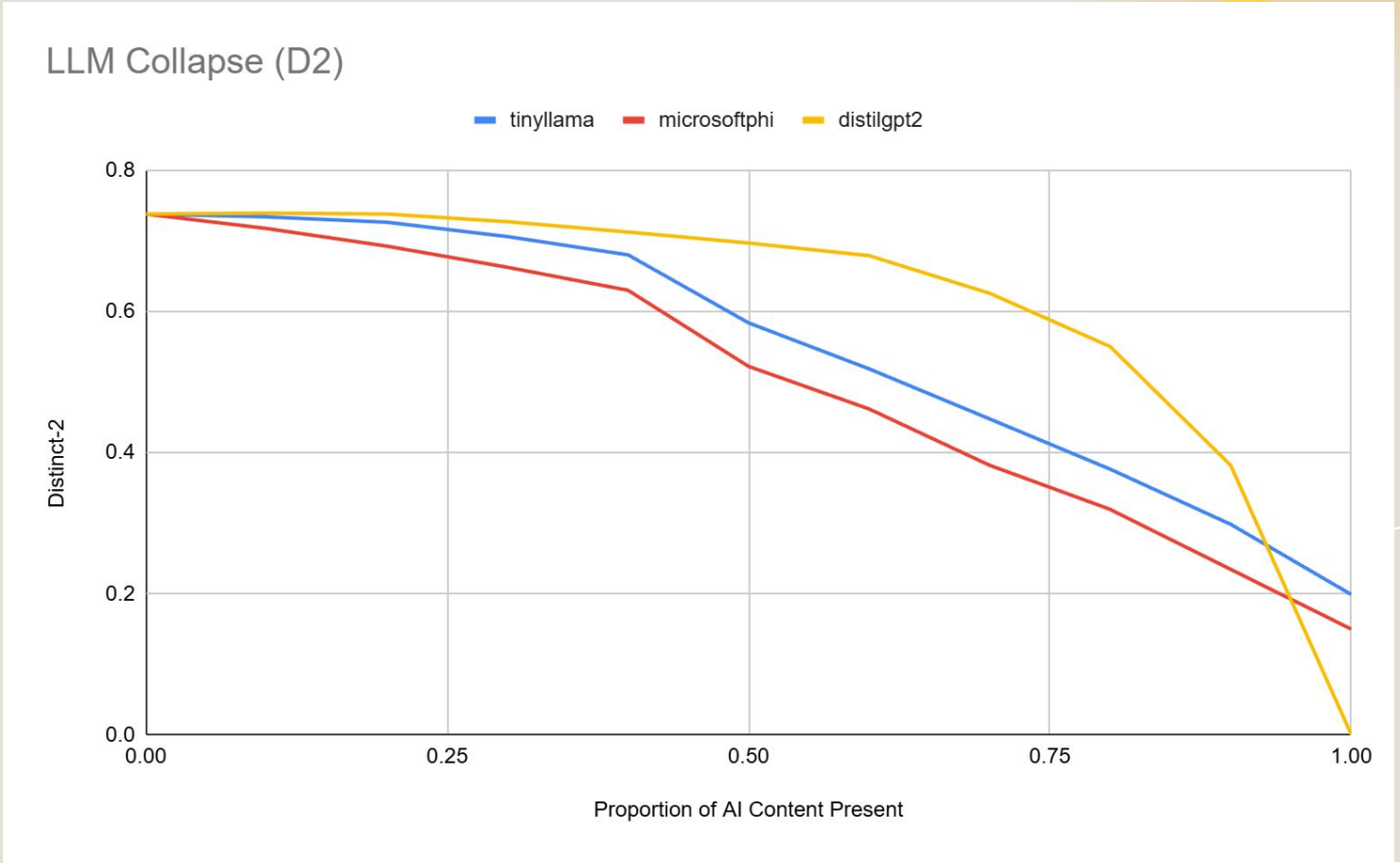
Simulation Results (1)



Simulation Results (2)

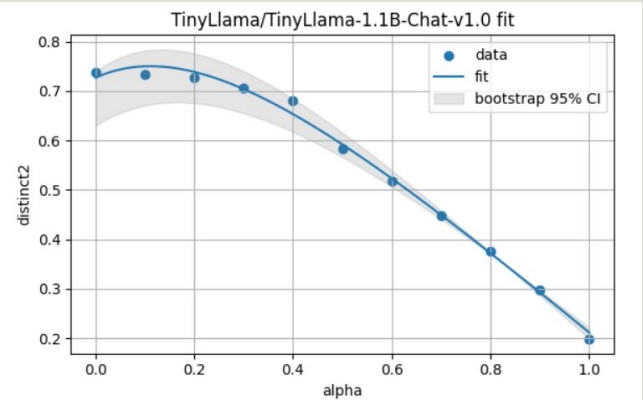


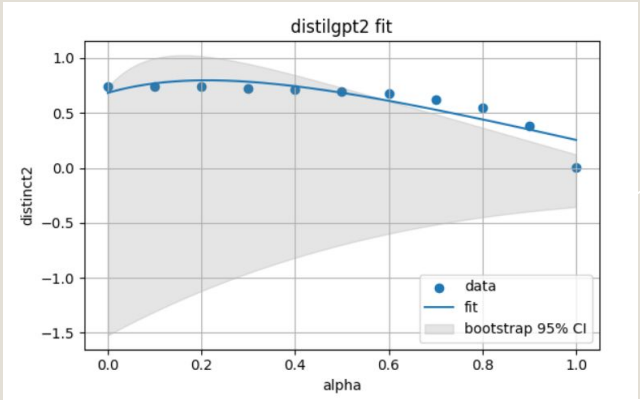
Simulation Results (3)

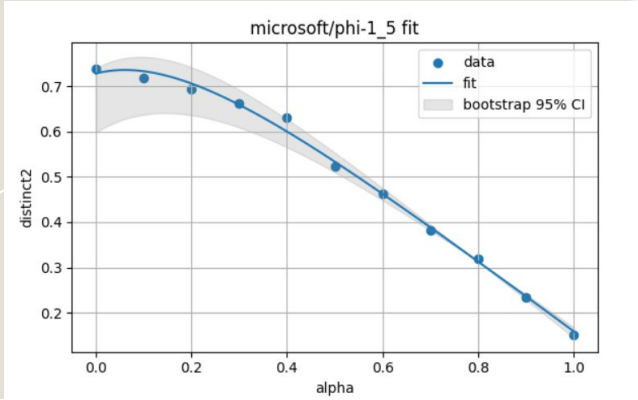


Fitting the Model

- Using a nonlinear least squares fit along with a bootstrap for confidence intervals
- Goal is for parameter estimation







Parameter Estimations

model	lam_est	k_est	Q0_est	lam_95lo	lam_95hi	k_95lo	k_95hi	Q0_95lo	Q0_95hi
TinyLlama/TinyLlama-1.1B-Chat-v1.0	3.2521	3.855	0.7268	2.8141	4.0332	3.4682	4.4572	0.63	0.7393
distilgpt2	3.8152	3.7722	0.6823	0.8326	12.1464	0.9655	9.9457	-1.5223	0.7358
microsoft/phi-1_5	3.832	4.8909	0.7278	3.2649	4.8929	4.3574	5.8028	0.5976	0.7391

MSE TinyLlama: 0.000141
MSE distilgpt2 : 0.009359
MSE phi-1_5 : 0.000157

Final Takeaways

- As α increases, the Distinct-2 diversity drops, confirming that recursive training leads to linguistic variety

- While different LLMs are affected differently, result is model decay regardless

- Our mathematical model fits observed patterns - suggesting that human content injection is essential to prevent LLM model collapse

References

- Shumailov, I., Shumaylov, Z., Zhao, Y. et al. *AI models collapse when trained on recursively generated data. Nature* **631**, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y

- Zhang, X., Zhou, H., & Palm, R. *“AG News: A text classification dataset.”* (2015).