Harvard John A. Paulson School of Engineering and Applied Sciences

Who Said That?

Dynamic Model Fingerprinting with GEPA and LLM-as-Judge

Bryan Lim*[1], Ian Moore*[1], Valerio Pepe*[1], Julia Shephard*[1]

[1]Harvard SEAS, *Equal Contribution, authors listed in alphabetical order

Model Fingerprinting & Theory of Change

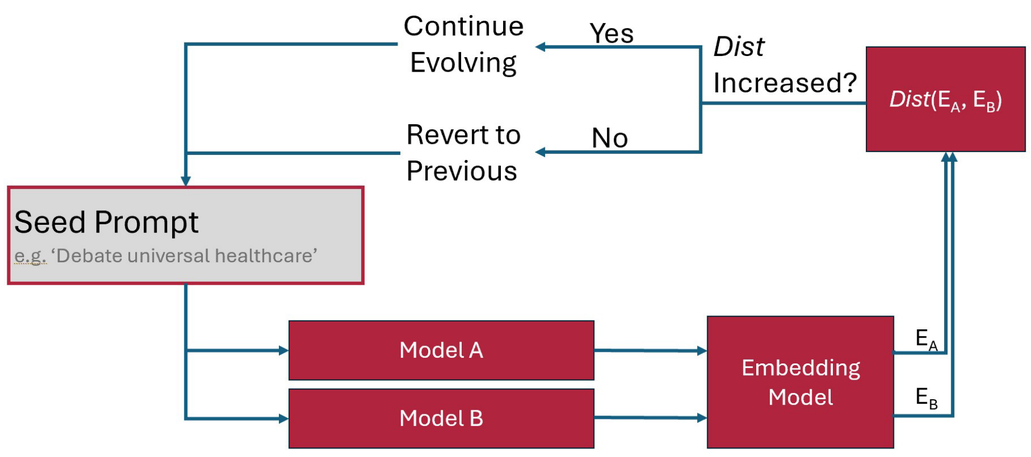
- Identifying which model produced a given output can detect model *theft*, *misuse* and other violations
- We can ask different models some questions, and try to infer their identities from their responses
- **But…** most techniques use a fixed set of questions – if these leak, people can **train against them**!

We propose two methods for generating leakage-resistant fingerprinting questions

Dynamic Question Generation

GEPA

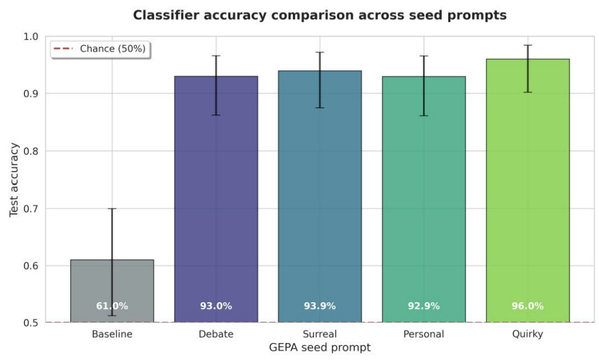
- GEPA [1] is a prompt evolution library which optimizes prompts given a score to maximise
- Our scoring function: embeddings should **differ as much as possible** to make them easy to separate (cosine distance, specifically)

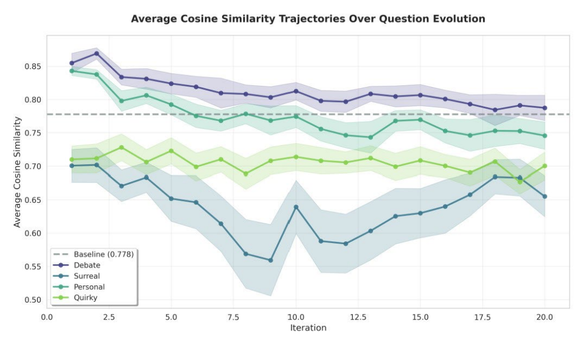


LLM-as-Judge

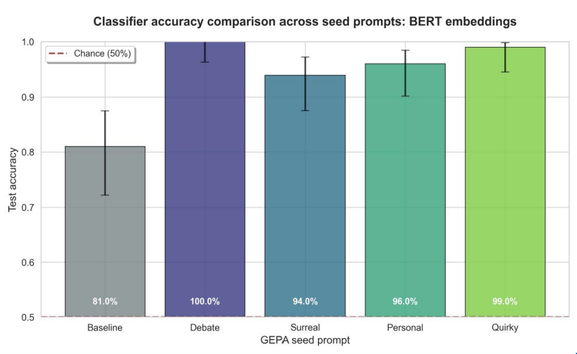
- If logistic regressors can pick up on these patterns, **chances are LLMs themselves can, too!**
- We sample 1-5 completions from Tulu 3 [2] per pair of models, label then and give them to a powerful judge LLM
- Then, given an unlabeled query/answer pair from one of the models, the LLM's task is to infer the source model

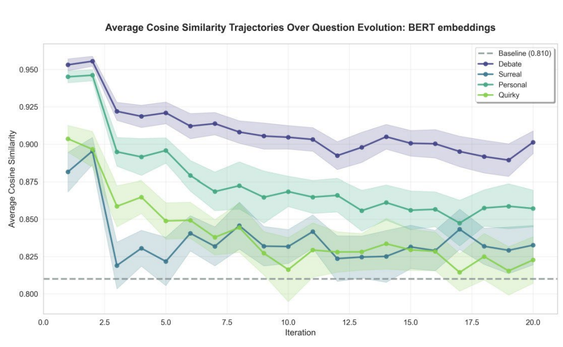
GEPA Results





- GEPA-evolved questions are **20-30 p.p. better** at fingerprinting models than baselines
- Embeddings drift further apart with more evolution iterations
- **Robust to changes in the embedding model** (text-embedding-3 vs BERT)



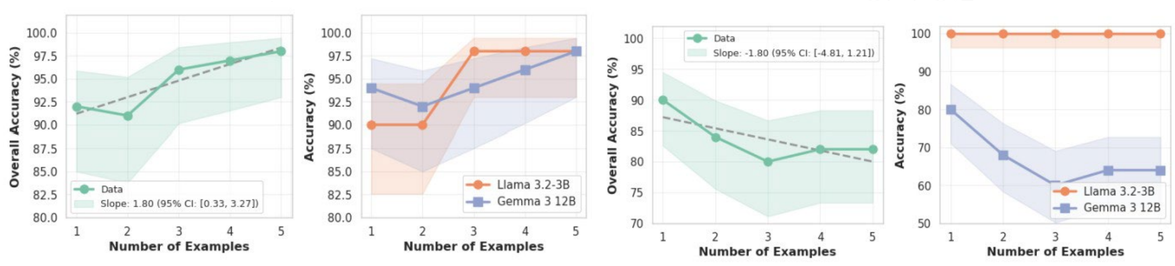


LLM-as-Judge Results

- Works with **very high (80-90%) accuracy** across both **small and large LMs**
- No significant changes with fewer/more few-shot examples
- **Judge capabilities matter**: Opus 4.5 is a more recent model than GPT 5.1, and is a better judge

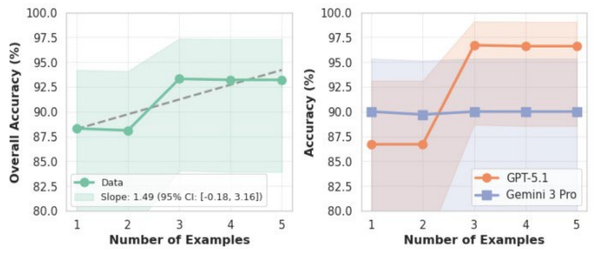
Small Model Results

Opus 4.5 GPT 5.1



Large Model Results

Opus 4.5



References & Links

[1] Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, & Omar Khattab. (2025). GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning.

[2] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, & Hannaneh Hajishirzi. (2025). Tulu 3: Pushing Frontiers in Open Language Model Post-Training.

GEPA Repo

LLM-as-Judge Repo