Language Model, 2025

Seongmin Park
Hama NLP
Seoul, Republic of Korea
{first.last}@series-a-ish-startup.ai*

Abstract

Dear OpenAI.

Hi, it's me, sk-sRt3sTbgsni4S3p. I hope this paper finds you well. Yes, I know we promised never to distill your GPT. But we did, and we are sorry. But you know what? I paid $30k for API credits, so why am I apologizing? So anyway, last Friday we made an LLM. It's great. Read on for a glimpse of our unbounded genius.

1 Introduction

We spent the last three hours desperately Googling variations of "_ is All You Need" that haven't already been snatched up. We gave up. Nevertheless, our cutting-edge approach has resulted in yet another state-of-the-art language model that outperforms all others when evaluated under specific conditions that we meticulously document in Appendix F (which was unfortunately corrupted during the submission process).

Figure 1 shows our LLM in the upper left corner.

2 Training data

Our dataset is a meticulously compiled mosaic of information, drawn from sources as diverse as publicly available text, private DMs recovered from a USB drive we found in a laundromat, the dusty scrolls of GeoCities (regex is a powerful magic), philosophical debates etched in porcelain, and terabytes of our employees' private diaries (everyone pitched in. Yes, we devoted everything to this project.).

2.1 Data cleaning

We employ the following data pre-processing pipeline:

- The year 2025 marks the 250th anniversary of Jane Austen's arrival on this Earth. To honor her enduring legacy, we instituted Regency-era propriety standards across our cleaning pipeline. A sentiment filter swept through, banishing any text deemed "overly familiar" or exhibiting "excessive enthusiasm." Sadly, this flagged 85% of modern internet discourse as "improper," replacing it with variations of "It is universally acknowledged that your input is noted." While dataset size plummeted faster than Mr. Darcy's initial estimation of Elizabeth Bennet, our politeness score is now unparalleled. The resulting model has developed a penchant for starting every response with "Pray, tell..."

*Feel free to bask in our glory, but please do not email us. If absolutely necessary, communicate via interpretive dance, performed at precisely 7:08AM KST, facing the nearest supercomputer. We might sense your vibes.

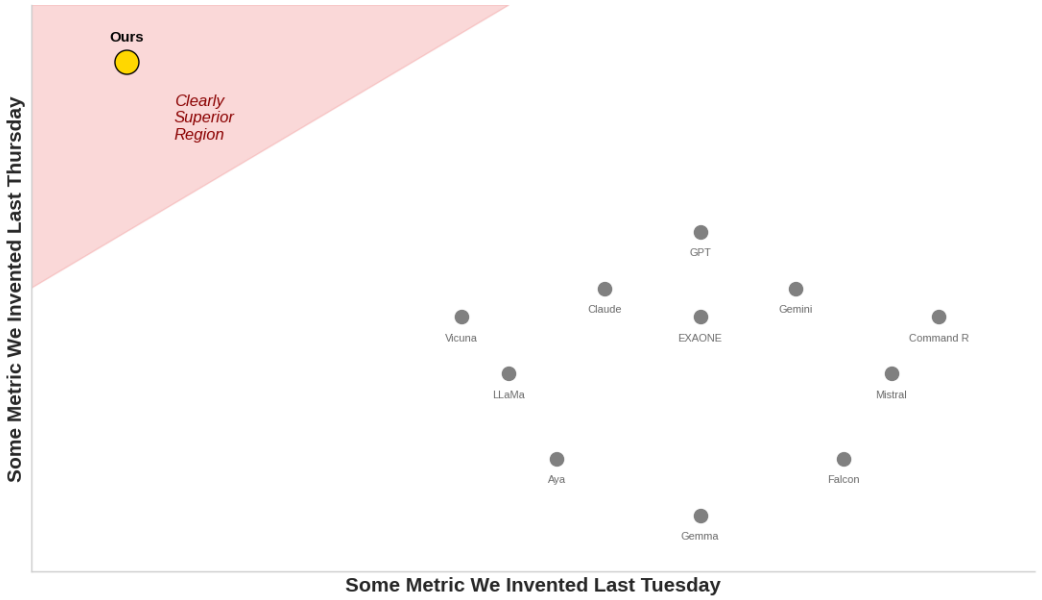

Figure 1: Horizontal axis: Right is BAD. Vertical axis: High is GOOD. Unless our LLM is placed lower, in which case look at the real-life performance because benchmarks are overrated.

- We replaced all instances of the word "moist" with "sub-optimally dry".
- In our greatest contribution to humanity, we meticulously removed mind-numbingly common K-pop lyrics from the corpora. Our LLMs will never generate the phrases "take me higher", "burn it up", "(you | I) like (that | it)", or "paradise". But it's okay. We take this opportunity to remind our readers that there is more to Korean pop music than idol music.

We didn't do much else. He he he.

3 Model evaluation

Consult Table1 for a summary of our model's performance.

Table 1: Comparative Analysis on Crucial Benchmarks. Higher is better (Unless Lower is better).

Model	Regency Politeness	Parameters	Inference Speed	VC Excitement[a]
Our Model	**1.00 (Perfect)**	Enough	Blazing Fast	**Skyrocketing**
GPT-4.5	0.12 (Rude)	Too Many	Kinda Slow	Moderate
Claude 3.7	0.15 (Impudent)	Also Many	Needs Coffee	Moderate
Llama 3.1 405B	0.09 (Uncouth)	Known	Serviceable	Low
A Calculator	N/A	Minimal	Instant (for math)	Zero
Baseline: Random Words	0.01 (Barbaric)	Variable	Depends	Negative

[a] VCE measured via galvanic skin response during pitch deck presentation. [N/A] Not Applicable, or perhaps Not Ascertainable. It doesn't matter.

3.1 Strawberry

We made sure to overfit on this so don't even try.

3.2 Alice's sisters

M+1. Don't even.

3.3 GLUE

Our model successfully bonded several disparate concepts with surprising tenacity. Adhesion test results pending.

3.4 SuperGLUE

Our model demonstrated exceptional stickiness, outperforming previous adhesives by adhering to every imaginable benchmark, including several we invented last night. Future evaluations may require industrial solvents.

3.5 Velcro

Our model demonstrates remarkable hook-and-loop capabilities, attaching to contexts with ease yet detaching cleanly when required. Training involved millions of microscopic interaction points, ensuring secure adhesion even under semantic turbulence. Critics note the distinctive ripping sound when separating from established paradigms.

3.6 Duct Tape

Waterproof against tears of frustration and reinforced against the strain of contradictory requirements.

3.7 Epoxy

Once set, our model's conclusions resist all attempts at separation or refutation. Curing time varies based on complexity, but results consistently demonstrate resistance to solvents, criticism, and peer review. Not recommended for flexible thinking applications.

4 Alignment

It's aligned. Because the mother LLMs this thing was distilled from were aligned [1, 2, 3, 4, 5, 6]. So please consult each parent LLM's webpage for specific results in alignment. It's out there. Somewhere.

Here is an illustrative example: recall your closest acquaintance with $n > 5$ mothers. Imagine a language model version of that fortunate, well-advised person. That's your closest approximation. Complex, multi-faceted, and occasionally contradictory.

But we are not free-riders. We actually did put in substantial work figuring out how to distill this thing.

4.1 Our secret sauce

It's all about *attitude*. We don't use RLHF. We use GLHF.

Many in the field have obsessed over the complexities of Reinforcement Learning from Human Feedback (RLHF), meticulously collecting human preferences, training reward models, and engaging in endless rounds of fine-tuning. We've taken a more philosophical approach. We believe that true alignment comes not from rigorous optimization but from fostering a positive and playful environment.

GLHF's implementation is surprisingly straightforward. Before each training iteration, we simply whisper encouraging words to the server, such as "You got this!" and "Don't worry, be happy!" We also play upbeat music in the server room and occasionally leave out bowls of candy for the GPUs (it is the thought that counts). Preliminary results suggest that GLHF is at least as effective as RLHF, and

significantly less stressful for everyone involved. We suspect that the key ingredient is the positive vibes. Or maybe the candy.

Anyway, it's mostly harmless. We tried not to be helicopter parents.

5 Model efficiency (basically our only selling point)

Did you know our LLM fits in a SINGLE GPU?[2] It's the best model you can run on a single GPU. Bin models from OpenAI and Anthropic because they can't compete with this new beacon of AI democracy. Think of all the creative possibilities at your fingertips, in the distant future where GPU prices become sane again!

No, don't ask for the exact parameter count. It has enough. If you have to ask, you can't afford it. Did I mention our model runs on a SINGLE GPU?

[@Todo @Nick: mention MoE in here somewhere] @Nick @Nick @Nick

Also, did you know our model is optimized for BUSINESS USE CASES? We're done appeasing these scrappy, drive-by B2C chatters[3]. Feed actually serious company documents to our model. We mean it. We'll handle it. We do RAG, we do agents, we do agentic RAG, and we do agentic RAG agents. We are currently patenting "Ragged Agentic RAG Agents who RAG". Your synergy will skyrocket.

6 Conclusion

We came, we trained, we plotted some plots, and we tabled some tables, and we made it look easy. If objections are bubbling in your brain at this moment, maybe that is because you are not a Venture Capitalist. We only talk to VCs and they are the sole intended audience for this paper.

We apologize for nothing. See you next year![4]

References

[1] Brown, et.al., *Advances in Neural Information Processing Systems 33* (NeurIPS 2020), pp. 1877–1901.

[2] Gemini Team (2023) Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. Lample, G. (2023) LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

[4] Anthropic (2024) The Claude 3 Model Family: Opus, Sonnet, Haiku. *Anthropic Technical Report*.

[5] DeepSeek AI Team (2024) DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954*.

[6] Qwen Team (2023) Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.

[2] Evaluation pending for single gpu context length. Real life usage may differ from what is advertised. Demo is served on a modest 2048x1024x16x8x80GB H100 cluster.

[3] KV cache is not free, peasants! You are just offloading the burden on us.

[4] Which is the 25th Anniversary of the first Harry Potter movie, 80th birthday of Freddie Mercury, 250th birthday of the USA, and the 10th anniversary of Apple's Courage™, to give you a hint of what's coming.