Google "We Have No Moat, And Neither Does OpenAI"

Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI

DYLAN PATEL AND AFZAL AHMAD

MAY 4, 2023 · PAID

653 10 Share

The text below is a very recent leaked document, which was shared by an anonymous individual on a public Discord server who has granted permission for its republication. It originates from a researcher within Google. We have verified its authenticity. The only modifications are formatting and removing links to internal web pages. The document is only the opinion of a Google employee, not the entire firm. We do not agree with what is written below, nor do other researchers we asked, but we will publish our opinions on this in a separate piece for subscribers. We simply are a vessel to share this document which raises some very interesting points.

We Have No Moat

And neither does OpenAI

We’ve done a lot of looking over our shoulders at OpenAI. Who will cross the next milestone? What will the next move be?

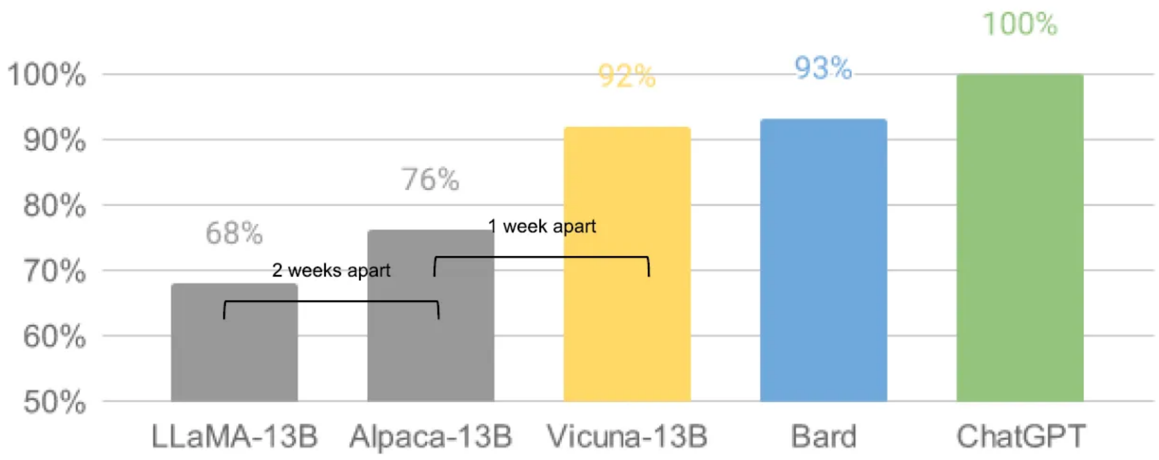
But the uncomfortable truth is, *we aren’t positioned to win this arms race and neither is OpenAI*. While we’ve been squabbling, a third faction has been quietly eating our lunch.

I’m talking, of course, about open source. Plainly put, they are lapping us. **Things we consider “major open problems” are solved and in people’s hands today.** Just to name a few:

- **LLMs on a Phone:** People are running foundation models on a Pixel 6 at 5 tokens / sec.
- **Scalable Personal AI:** You can finetune a personalized AI on your laptop in an evening.
- **Responsible Release:** This one isn't "solved" so much as "obviated". There are entire websites full of art models with no restrictions whatsoever, and text is not far behind.
- **Multimodality:** The current multimodal ScienceQA SOTA was trained in an hour.

While our models still hold a slight edge in terms of quality, the gap is closing astonishingly quickly. Open-source models are faster, more customizable, more private, and pound-for-pound more capable. They are doing things with $100 and 13B params that we struggle with at $10M and 540B. And they are doing so in weeks, not months. This has profound implications for us:

- **We have no secret sauce.** Our best hope is to learn from and collaborate with what others are doing outside Google. We should prioritize enabling 3P integrations.
- **People will not pay for a restricted model when free, unrestricted alternatives are comparable in quality.** We should consider where our value add really is.
- **Giant models are slowing us down.** In the long run, the best models are the ones which can be iterated upon quickly. We should make small variants more than an afterthought, now that we know what is possible in the <20B parameter regime.



*GPT-4 grades LLM outputs. Source: https://vicuna.lmsys.org/

https://lmsys.org/blog/2023-03-30-vicuna/

What Happened

At the beginning of March the open source community got their hands on their first really capable foundation model, as Meta's LLaMA was leaked to the public. It had no instruction or conversation tuning, and no RLHF. Nonetheless, the community immediately understood the significance of what they had been given.

A tremendous outpouring of innovation followed, with just days between major developments (see The Timeline for the full breakdown). Here we are, barely a month later, and there are variants with instruction tuning, quantization, quality improvements, human evals, multimodality, RLHF, etc. etc. many of which build on each other.

Most importantly, they have solved the scaling problem to the extent that anyone can tinker. Many of the new ideas are from ordinary people. The barrier to entry for training and experimentation has dropped from the total output of a major research organization to one person, an evening, and a beefy laptop.

Why We Could Have Seen It Coming

In many ways, this shouldn't be a surprise to anyone. The current renaissance in open source LLMs comes hot on the heels of a renaissance in image generation. The similarities are not lost on the community, with many calling this the "Stable Diffusion moment" for LLMs.

In both cases, low-cost public involvement was enabled by a vastly cheaper mechanism for fine tuning called low rank adaptation, or LoRA, combined with a significant breakthrough in scale (latent diffusion for image synthesis, Chinchilla for LLMs). In both cases, access to a sufficiently high-quality model kicked off a flurry of ideas and iteration from individuals and institutions around the world. In both cases, this quickly outpaced the large players.

These contributions were pivotal in the image generation space, setting Stable Diffusion on a different path from Dall-E. Having an open model led to product integrations, marketplaces, user interfaces, and innovations that didn't happen for Dall-E.

The effect was palpable: rapid domination in terms of cultural impact vs the OpenAI solution, which became increasingly irrelevant. Whether the same thing will happen for LLMs remains to be seen, but the broad structural elements are the same.

What We Missed

The innovations that powered open source's recent successes directly solve problems we're still struggling with. Paying more attention to their work could help us to avoid reinventing the wheel.

LoRA is an incredibly powerful technique we should probably be paying more attention to

LoRA works by representing model updates as low-rank factorizations, which reduces the size of the update matrices by a factor of up to several thousand. This allows model fine-tuning at a fraction of the cost and time. Being able to personalize a language model in a few hours on consumer hardware is a big deal, *particularly* for aspirations that involve incorporating new and diverse knowledge in near real-time. The fact that this technology exists is underexploited inside Google, even though it directly impacts some of our most ambitious projects.

Retraining models from scratch is the hard path

Part of what makes LoRA so effective is that - like other forms of fine-tuning - it's stackable. Improvements like instruction tuning can be applied and then leveraged as other contributors add on dialogue, or reasoning, or tool use. While the individual fine tunings are low rank, their sum need not be, allowing full-rank updates to the model to accumulate over time.

This means that as new and better datasets and tasks become available, the model can be cheaply kept up to date, without ever having to pay the cost of a full run.

By contrast, training giant models from scratch not only throws away the pretraining, but also any iterative improvements that have been made on top. In the open source world, it doesn't take long before these improvements dominate, making a full retrain extremely costly.

We should be thoughtful about whether each new application or idea really needs a whole new model. If we really do have major architectural improvements that preclude directly reusing model weights, then we should invest in more aggressive forms of distillation that allow us to retain as much of the previous generation's capabilities as possible.

Large models aren't more capable in the long run if we can iterate faster on small models

LoRA updates are very cheap to produce (~$100) for the most popular model sizes. This means that almost anyone with an idea can generate one and distribute it. Training times under a day are the norm. At that pace, it doesn't take long before the cumulative effect of all of these fine-tunings overcomes starting off at a size disadvantage. Indeed, in terms of engineer-hours, the pace of improvement from these models vastly outstrips what we can do with our largest variants, and the best are already largely indistinguishable from ChatGPT. **Focusing on maintaining some of the largest models on the planet actually puts us at a disadvantage.**

Data quality scales better than data size

Many of these projects are saving time by training on small, highly curated datasets. This suggests there is some flexibility in data scaling laws. The existence of such datasets follows from the line of thinking in Data Doesn't Do What You Think, and they are rapidly becoming the standard way to do training outside Google. These datasets are built using synthetic methods (e.g. filtering the best responses from an existing model) and scavenging from other projects, neither of which is dominant at Google. **Fortunately, these high quality datasets are open source, so they are free to use.**

Directly Competing With Open Source Is a Losing Proposition

This recent progress has direct, immediate implications for our business strategy. **Who would pay for a Google product with usage restrictions if there is a free, high quality alternative without them?**

And we should not expect to be able to catch up. The modern internet runs on open source for a reason. Open source has some significant advantages that we cannot replicate.

We need them more than they need us

Keeping our technology secret was always a tenuous proposition. Google researchers are leaving for other companies on a regular cadence, so we can assume they know everything we know, and will continue to for as long as that pipeline is open.

But holding on to a competitive advantage in technology becomes even harder now that cutting edge research in LLMs is affordable. Research institutions all over the world are building on each other's work, exploring the solution space in a breadth-first way that far outstrips our own capacity. We can try to hold tightly to our secrets while outside innovation dilutes their value, or we can try to learn from each other.

Individuals are not constrained by licenses to the same degree as corporations

Much of this innovation is happening on top of the leaked model weights from Meta. While this will inevitably change as truly open models get better, the point is that they don’t have to wait. The legal cover afforded by “personal use” and the impracticality of prosecuting individuals means that individuals are getting access to these technologies while they are hot.

Being your own customer means you understand the use case

Browsing through the models that people are creating in the image generation space, there is a vast outpouring of creativity, from anime generators to HDR landscapes. These models are used and created by people who are deeply immersed in their particular subgenre, lending a depth of knowledge and empathy we cannot hope to match.

Owning the Ecosystem: Letting Open Source Work for Us

Paradoxically, the one clear winner in all of this is Meta. Because the leaked model was theirs, they have effectively garnered an entire planet's worth of free labor. Since most open source innovation is happening on top of their architecture, there is nothing stopping them from directly incorporating it into their products.

The value of owning the ecosystem cannot be overstated. Google itself has successfully used this paradigm in its open source offerings, like Chrome and Android. By owning the platform where innovation happens, Google cements itself as a thought leader and direction-setter, earning the ability to shape the narrative on ideas that are larger than itself.

The more tightly we control our models, the more attractive we make open alternatives. Google and OpenAI have both gravitated defensively toward release patterns that allow them to retain tight control over how their models are used. But this control is a fiction. Anyone seeking to use LLMs for unsanctioned purposes can simply take their pick of the freely available models.

Google should establish itself a leader in the open source community, taking the lead by cooperating with, rather than ignoring, the broader conversation. This probably means taking some uncomfortable steps, like publishing the model weights for small ULM variants. This necessarily means relinquishing some control over our models. But this compromise is inevitable. We cannot hope to both drive innovation and control it.

Epilogue: What about OpenAI?

All this talk of open source can feel unfair given OpenAI's current closed policy. Why do we have to share, if they won't? But the fact of the matter is, we are already sharing everything with them in the form of the steady flow of poached senior researchers. Until we stem that tide, secrecy is a moot point.

And in the end, *OpenAI doesn't matter*. They are making the same mistakes we are in their posture relative to open source, and their ability to maintain an edge is necessarily in question. Open source alternatives can and will eventually eclipse them unless they change their stance. In this respect, at least, we can make the first move.

The Timeline

Feb 24, 2023 - LLaMA is Launched

Meta launches LLaMA, open sourcing the code, but not the weights. At this point, LLaMA is not instruction or conversation tuned. Like many current models, it is a relatively small model (available at 7B, 13B, 33B, and 65B parameters) that has been trained for a relatively large amount of time, and is therefore quite capable relative to its size.

March 3, 2023 - The Inevitable Happens

Within a week, LLaMA is leaked to the public. The impact on the community cannot be overstated. Existing licenses prevent it from being used for commercial purposes, but suddenly anyone is able to experiment. From this point forward, innovations come hard and fast.

March 12, 2023 - Language models on a Toaster

A little over a week later, Artem Andreenko gets the model working on a Raspberry Pi. At this point the model runs too slowly to be practical because the weights must be paged in and out of memory. Nonetheless, this sets the stage for an onslaught of minification efforts.

March 13, 2023 - Fine Tuning on a Laptop

The next day, Stanford releases Alpaca, which adds instruction tuning to LLaMA. More important than the actual weights, however, was Eric Wang's alpaca-lora repo, which used low rank fine-tuning to do this training "within hours on a single RTX 4090".

Suddenly, anyone could fine-tune the model to do anything, kicking off a race to the bottom on low-budget fine-tuning projects. Papers proudly describe their total spend of a few hundred dollars. What's more, the low rank updates can be distributed easily and separately from the original weights, making them independent of the original license from Meta. Anyone can share and apply them.

March 18, 2023 - Now It's Fast

Georgi Gerganov uses 4 bit quantization to run LLaMA on a MacBook CPU. It is the first "no GPU" solution that is fast enough to be practical.

March 19, 2023 - A 13B model achieves "parity" with Bard

The next day, a cross-university collaboration releases Vicuna, and uses GPT-4-powered eval to provide qualitative comparisons of model outputs. While the evaluation method is suspect, the model is materially better than earlier variants. **Training Cost: $300.**

Notably, they were able to use data from ChatGPT while circumventing restrictions on its API - They simply sampled examples of "impressive" ChatGPT dialogue posted on sites like ShareGPT.

March 25, 2023 - Choose Your Own Model

Nomic creates GPT4All, which is both a model and, more importantly, an ecosystem. For the first time, we see models (including Vicuna) being gathered together in one place. **Training Cost: $100.**

March 28, 2023 - Open Source GPT-3

Cerebras (not to be confused with our own Cerebra) trains the GPT-3 architecture using the optimal compute schedule implied by Chinchilla, and the optimal scaling implied by μ-parameterization. This outperforms existing GPT-3 clones by a wide margin, and represents the first confirmed use of μ-parameterization "in the wild". These models are trained from scratch, meaning the community is no longer dependent on LLaMA.

March 28, 2023 - Multimodal Training in One Hour

Using a novel Parameter Efficient Fine Tuning (PEFT) technique, LLaMA-Adapter introduces instruction tuning and multimodality in one hour of training. Impressively, they do so with just 1.2M learnable parameters. The model achieves a new SOTA on multimodal ScienceQA.

April 3, 2023 - Real Humans Can't Tell the Difference Between a 13B Open Model and ChatGPT

Berkeley launches Koala, a dialogue model trained entirely using freely available data.

They take the crucial step of measuring real human preferences between their model and ChatGPT. While ChatGPT still holds a slight edge, more than 50% of the time users either prefer Koala or have no preference. **Training Cost: $100.**

April 15, 2023 - Open Source RLHF at ChatGPT Levels

Open Assistant launches a model and, more importantly, a dataset for Alignment via RLHF. Their model is close (48.3% vs. 51.7%) to ChatGPT in terms of human preference. In addition to LLaMA, they show that this dataset can be applied to Pythia-12B, giving people the option to use a fully open stack to run the model. Moreover, because the dataset is publicly available, it takes RLHF from unachievable to cheap and easy for small experimenters.