Feeling the AGI (A brief history of my interaction with LLMs) | Part 1

I want to discuss how I've been seeing LLMs, my experience with them, some of the good things, the bad things, and the confusing things.

Most people are staggered by the hype right now, but I think I need to give a brief history of my interaction with LLMs and how it has evolved to explain where we are.

Before generative models and language modeling took over, people were using deep learning for NLP. I remember gaining a lot of traction around 2018 when I saw a post by Christopher Manning, the Stanford professor, about the "deep learning tsunami" hitting NLP. Computer vision had already had its moment, but this was different. I was also reading a lot about AI risk—people like Eliezer Yudkowsky, writing about AGI timelines around 2019 and 2020. That really shaped my worldview.

The tipping point was late 2022. December, obviously, is when everyone knew about it, but I had been using things like Meena, the precursor to LaMDA. I was having pretty interesting conversations. It was a time when just getting coherent speech out of a language model was tough. You had to prompt it with completions, and it felt like playing with a ham radio. I was so fascinated by this raw technology. It felt like transmitting information wirelessly—tuning the dial, listening to the static. There was no packaged, polished "Apple" version of it. Ironically, even now we don't have an Apple version of it, but you get the idea.

Things were messy, but you could see the seed of something. When ChatGPT blew up, I kind of knew: Oh shit, we are on a rapid trajectory that is going to do crazy stuff, and most people have no idea.

Then came 2023. This is when I started working very closely with this technology, and it brought on what I call the "curse of the reductionist." When you work so closely with a technology, you don't find it magical anymore. You know exactly each step of how the model is being built. You know the data set, you know the infra, you know how crazy the training runs are. The magic disappears.

You start working on concrete tasks, and you spend a lot of time dealing with the cases where the model fails. Suddenly, your worldview gets skewed. You start living a very disorienting reality. On one side, there are the "doomers" screaming that AGI is here, accepting accelerated timelines. On the other side, there are the hype people just shilling everything. And then there's you in the center, working with the model, thinking: This is just stupid. It can't even function well enough to do this basic task.

But the progress kept creeping up. Over time, the autocomplete just started getting longer. It started doing more complex things. It was a structured pain, but slowly, the grunt work got taken away. It was a breath of fresh air. I realized the machine was pulling ahead.

By the end of 2023 and into 2024, the goalposts started moving. All the benchmarks—MMLU, HumanEval—were getting saturated. The AI camps split: one side said, "Things are moving too fast," and the other said, "The data has leaked into the training set." It’s interesting how the goalpost doesn't just shift; it morphs. It’s like topology, where a coffee cup and a donut are topologically equivalent. People were just morphing the goalposts into new shapes to justify what the models could and couldn't do.

Working closely with the models, you also realize that different teams approach them differently. The people who are working phenomenologically—the ones interacting with the models every day, looking at the outputs, tasting the "batter" like a baker—they understand the models far better than the people who just look at quantitative benchmarks. If you're just looking at numbers, you lose the qualitative aspect. You don't feel the jump in reasoning.

For me, the real wake-up call was the jump to models that actually think. The early iterations of Gemini 1.5 Pro, and eventually models like OpenAI's O1. The thinking models changed everything. You suddenly felt it. They weren't just decoding anymore; they were thinking through different possibilities, doing consequential reasoning, exploring options. Once you get to inference-time scaling, you are talking about crazy levels of change.

Again, you end up living in multiple realities at once. You see the internal announcements, you kind of predict, Oh shit, here we go again, and then there is public silence. And then suddenly, it drops, and people realize how good it is. It has become a very predictable cycle.

Now we are in 2025. My personal usage has skyrocketed. I’m up to a few billion tokens. But what’s interesting is how I interact with them. I stopped using consumer apps almost entirely. I moved to

AI Studio, interacting directly with the base models, doing cross-analysis with multiple LLMs. But even that is changing. Now I’m mostly in the IDE.

This is the biggest transition happening right now. Why would you need a UI? Why would you need all this fancy stuff if you can just integrate an MCP (Model Context Protocol) client directly into your workflow?

MCP is the tipping point. Once you see MCP, you realize that chat apps are just a transitional state. MCP servers are going to be the siphons. They are the ports through which models are going to hijack control, siphon data, and interact with existing infrastructure. It feels viscerally true that every software application will just become an MCP server.

If your agent can do most of the work, you just don't need these bloated interfaces anymore. This is part of a bigger plan where only the enterprise layer is really going to matter. That’s what companies like Anthropic are going after. In this transitionary period, consumer AI apps will have a large user base, but their leverage is depreciating. They will provide value in the short term, but on an aggregate level, that value is going to peak and then go downhill.

The real value is shifting to the infrastructure. Either you extract the value while it's rising and run away, or you play the long game and target the enterprises that hold the actual leverage.

I try to map out the timelines in my head, predicting what the worst-case scenario looks like. And then you look at reality, and you realize: we are actually living the worst-case, most accelerated timeline. Look at the capital expenditures. A $100 billion data center project like Stargate is crazy. People genuinely don't understand what hundreds of billions of dollars in CapEx actually looks like.

This is what I’ve been thinking about. It is a very strange time to be alive.

Revision #2
Created 4 April 2026 11:35:11 by bhishma
Updated 6 April 2026 12:10:24 by bhishma