The First Act

In 2023, ChatGPT was released to the world, fundamentally changing how technology is created and how we think about businesses. What once took weeks to build now takes hours to minutes with a simple prompt and a few clicks. Tools like Cursor and Codex brought AI to our fingertips, allowing us to seamlessly interact with AI on a daily basis. Businesses and individuals claim the "improved productivity gains" that AI offers while consumers enjoy an endless stream of content and information delivered to their pocket sized pieces of glass. In society, we are seeing a rapid shift in the norms of AI in everyday life as well: churches are using AI to interact with a virtual "Jesus"[1], schools are debating with the implications where learning is no longer a linear process, and governments are scrambling to regulate a technology they barely understand. All of this being underlined by the "arms race" between AI Labs as we race towards so-called superintelligence.

However, we do not aim to discuss how AI will affect the world in this essay, but instead, to explore upcoming trends and where AI will go in the next 6 months. For those who want to see the predictions, just scroll to the callout boxes scattered through the article.

AI's Second Act

The First Act of AI was about building the foundational models and infrastructure that brought AI to a wide audience. My predictions for the Second Act of AI is that we will see a rapid commodification of AI models and applications. And so begins the Second Act.

I: Model Performance Converges to Parity

I personally don't think that there is much difference between the performance of models from OpenAI, Anthropic, and other labs. Sure, maybe there are marginal improvements of 1-2% on existing benchmarks, but in reality, it is barely noticeable to the average user. In 2023-2024, this was a major selling point for differentiation between others, but it doesn't seem to be the case now in 2025 and onwards. Additionally, a whole new host of open-source models that are either research projects or released from frontier labs which match much of the performance the leading labs claim but at a fraction of the cost. Some open source models you should check out are: AI2's new

OLMO models, Arcee's Trinity, and Google's Gemma models.

> **Prediction:** We will see a new breed of models come up where the architecture or components of it are specifically tailored to a domain or use-case. Examples of these use cases are developing specific architectures for domains such as electronic health records, weather time series, or financial data.
> **Counterpoint:** The bitter lesson will strike that general purpose models will dominate any vertical engineering research and development. Richard Sutton's *The Bitter Lesson* states that general approaches to compute will always beat out specific approaches[2].
> **Prediction:** We will see an increase in "personalized AI-first devices" where the hardware is designed specifically to run small open-source models locally as users become more concerned about privacy. Perhaps even as models become intelligent enough at an efficient compute-perspective, there will be a widespread trend of local-first AI.

II: GPUs Get Easier

Compute Unified Device Architecture, better known as CUDA, is a parallel computing platform and model developed by Nvidia which allows developers to utilize Nvidia GPU chips for general purpose and parallel computing. Since its release, CUDA has become the standard for AI training and inference workloads[3]. However, if you look at every major hyperscalar, they have some version of their own chips. Nvidia has their H-series, Google with TPUs, Amazon with Trainium, and Microsoft with AMD. From this pattern, it seems that each large player is positioning themselves to start vertically integrating themselves from the chip and inference to the model level.

While it seems that Nvidia's chips have been dominating news headlines, Google has been slowly gaining ground on their TPU chips. In fact, the newest model, Gemini 3.0 Pro was trained only on TPUs[4] while Anthropic and Google signed a deal for Anthropic to utilize up to 1 million TPUs for future model training and inference[5] Additionally, Microsoft is conducting research to convert CUDA to AMD-native workloads[6] while Google is doing something similar to convert to TPU-native workloads[7]. With tools that can broadly transition to other stacks, what is the benefit of CUDA architecture?

> **Prediction:** Nvidia will not have that strong of a moat due to advances in other chips such as TPUs and AMD chips. We will see a future where workloads and model performance are not specifically reliant on a brand of chips.
> **Counterpoint:** These workload-conversion tools will lack full integration/transformation between different architectures. This means transitioning between different architectures will take months of development and rewrite for potentially marginal gain. Nvidia's deep integration, developer experience, and continuously improving hardware will continue holding the lines in the near term. Additionally, expansion into robotics, space, and other domains should not be brushed off lightly.

III: Network Effects, Brand Power, and Behavior

As people begin to use more and more AI tooling, the question becomes, what is the "cool" tool to use. For example, Anthropic's series of Claude models are synonymous with coding, Nano Banana by Deepmind is known for visual generation, and OpenAI's GPT series is known for just general purpose usage. As these models improve over time, they will only reinforce some of the network effects that the foundational labs have built. Combine that with the expansion and unlimited resources of these labs, **we will see a quick crush of new AI application startups that will natively be added to the model offerings.**

> **Prediction:** Vertical AI applications will thrive more than ever in a growing world of general-purpose approaches. Labs do not have the resources to build everything and anything, moreso develop the UX and domain expertise to build vertical applications without losing focus on their core models.
> **Counterpoint:** However, this does not mean that the bitter lesson will strike in the future and that general purpose models will dominate any semblance of vertical engineering work. Richard Sutton's *The Bitter Lesson* states that general approaches to compute will always beat out specific approaches.

This does not mean that all hope is lost for new technology startups. I still believe that there is still room to grow. Applications are a user's first interaction with the technology itself that powers it. When we look at who we are building for, it is the user. Unlike developers, engineers, or other people within our space, when you look

outside, people are not living in developer tools, dashboards, or evaluations.[8] There are opportunities to build *experiences* instead of applications.

> **Prediction:** New ventures that own every aspect of the user's experience from creating meaningful intent, feedback, and embed themselves as deep as possible into their workflows will thrive. In this case, collecting as much data that inform behavior, patterns, value created, and intent will allow you to amass:
>
> 1) Unique datasets that are impossible to replicate by frontier labs.
>
> 2) User loyalty, behavior, and brand power that will allow users to stick with your product even if a better alternative comes along.

IV: Physical AI

Physical AI is one topic I am extremely interested in seeing how the landscape will evolve in the next 5 years. Currently, we are at a tipping point. Only up until 3 years ago, robotics were barely usable and the internet found amusement in watching researchers knock over and purposely trip robots[9]. Since then, we have seen an almost exponential growth of robotic capabilities over time. In 2024, Figure raised $675 million dollars for humanoid development[10], Boston Dynamics came out with an all-new Atlas[10], and an increasing number of laboratories dedicated to building robotic world-models have come up since then. In 2025, that trend continued with new startup SkildAI raising at a mega $4.5 billion valuation[11], UniTree's crazy-agile demos of their robots[12] and more are only continuing the trend. However, much of these demonstrations of technology have still been in a very sandboxed environment. That is until 1x.Tech's Neo robotics presented itself as a $500/month subscription service to bring a robot into the house with a very compelling demo[13]. The caveat (if you look at the fine text of Neo), is that the product itself is ***teleoperated*** meaning that you are paying another person to operate and have a window into your home through a robot. While it is the source of entertaining memes and ridicule, I believe that what we are seeing now is only the beginning.

> **Prediction**: Robots will have their "ChatGPT-moment" in the near-future. Just like how in 2022 when ChatGPT released and spurred a revolution in model performance, speed, and accuracy, robotics will experience something similar too. I

> believe that once a sub-$2,000 robot with real functionality is attained, we will see an explosion of use-cases and performance.
> **Counterpoint:** Technology and supply chain limitations in components such as actuators, batteries, and more may prevent physical AI from meaningfully integrating within society. However, that doesn't mean sectors such as manufacturing and transportation won't benefit from such advances.

V: Automated Labs

Over the past 3 years, it feels like we have focused more on building ChatGPT-wrappers that launch with a flash-in-the-pan rather than the true deep sciences. Scrolling through Hacker News, it feels like a never-ending carousel with "Cursor for X, Cursor for Y, or something similar." Launching a startup meant going to v0, spinning up a landing page, putting a UI in front of an OpenAI API, and calling it a product. In fact, it almost felt magical[14]. However, those times are now over.

The current SoTA models are now at an inflection point. They have been good enough to reach what many people call "genius." For example, most SoTA models rank PhD-level performance in many domains and Gemini 3.0 Pro even reached an incredible 37.5% on Humanity's Last Exam (HLE)[4]. Now, I believe that the next frontier is utilizing AI to actively advance the natural sciences: mathematics, physics, biology, chemistry, and thermodynamics. Similarly, the reverse will hold true where we will see applications of AI in those fields.

Before, scientific experiments were constrained by human output. We could only perform experiments and conduct analysis as fast as the number of humans (and amount of funding dollars) we could throw at a problem. However, with AI now, we are now constrained by how many tokens we can output and if our experiment can really fit into a context window. Groups working on this problem include Periodic Labs[15] aiming to discover new superconductors, Axiom[16] trying to develop the first mathematical super-reasoner, and Deepmind's Co-Scientist[17] driving forward automated science and discovery are all early examples of such products.

> **Prediction**: We will see steady growth in the number of startups and ventures dedicated to driving forward scientific progress in the natural sciences. Pitches such as "AI-researcher for rare-earth metals alternatives" or "mathematical verification of cyberspace systems" will become more and more common.

> **Counterpoint:** The time horizon for investments in deep tech is much longer than traditional venture cycles. Additionally, these ventures require significant technology innovation, longer product development timelines, and most importantly, potential regulatory barriers and customer education before widespread adoption.

VI: Personalized Software

Right now, AI usage is only at a fraction of what it is possible in the near future. Although we may think that everyone uses AI in our immediate circles and it is all the rage, the reality is that only a fraction of the population uses it on a frequent basis. Pew Research studies state found that only 34% of Americans have used ChatGPT with a majority of adults under 30 having used it actively.[18]

In a more personal anecdote, speaking with friends and family members back home (Appalachia), many have heard of it but few have actually used it. Maybe some entertain image generation tools or a never-ending video scroll, or others use it to help draft an email, but the use is still not widespread. Many of them have cited reasons for it just not being useful to them in their daily lives or occupations. However, my belief is that this is only the beginning of a much broader movement as AI shifts towards it's Second Act.

> **Prediction:** For the next few months, I predict we will see a movement towards more personalized software that understands user intent and behavior on a much deeper level. Right now, many of these tools are still scattered. Sure, progress has been made with tools such as ChatGPT plugins and Claude Artifacts, but we are still a bit away from some form of stateful, personalized software that truly understands and evolves around the needs of the user. Who knows what a final product would look like? Is it a bundle of a coding agent, file store, front ends, or something even more complex? An interesting company working in this space is Zo which is working on intelligent personal software.

VII: Fusion Comes Back

Now who said we were only limited to software/hardware topics? After all, what will power all the ideas we just discussed? As of October 2025, AI data centers are set to consume 1,600 terawatt-hours of power demand by 2035, equal to 4.4% of global electricity[19]. Current data center construction is on a 5-15 year time scale which means

current build outs are not keeping up with the energy requirements of AI right now. Without getting into the economics piece, that raises the question of where will we get that energy?

> **Prediction:** We will see a resurgence in nuclear and fusion capabilities in the coming years. Despite the partial-meltdown of Three Mile Island in 1979, $1 billion was recently loaned to restart the plant to power Microsoft[20]. Additionally, new partnerships with the UAE[21] and Westinghouse[22] continue to add momentum. Power still remains to this day one of the largest bottlenecks in AI.
> **Counterpoint:** Efforts to restart and build out nuclear initiatives will encounter regulatory and political hurdles which will prevent the anticipated speed of nuclear adoption.

Conclusion

While we have seen remarkable progress over the past 3 years around AI, I believe that the next few years will be a time for even more rapid expansion as we start fully developing out applications for these technologies. From automated factories to advancing the actual sciences, we are in a time filled with opportunities to innovate. With that being said, so begins AI's Second Act.

Works Cited

[1] URL: https://x.com/remarks/status/1989446280716927298?s=20.

[2] URL: http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

[3] URL: https://medium.com/@aidanpak/the-cuda-advantage-how-nvidia-came-to-dominate-ai-and-the-role-of-gpu-memory-in-large-scale-model-e0cdb98a14a0.

[4] URL: https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

[5] URL: https://www.cnbc.com/2025/10/23/anthropic-google-cloud-deal-tpu.html.

[6] URL: https://wccftech.com/microsoft-has-developed-toolkits-to-break-nvidia-cuda-dominance/.

[7] URL: https://x.com/RihardJarc/status/1989356949591204313.

[8] URL: https://x.com/natashamalpani/status/1987976089134899642?s=20.

[9] URL: https://www.youtube.com/watch?v=aX7KypGlitg.

[10] URL: https://spectrum.ieee.org/top-robotics-stories-2024.

[11] URL: https://finance.yahoo.com/news/nvidia-samsung-back-4-5b-203049332.html.

[12] URL: https://www.unitree.com.

[13] URL: https://www.1x.tech.

[14] URL: https://dev.to/dev_tips/the-graveyard-of-ai-startups-startups-that-forgot-to-build-real-value-5ad9.

[15] URL: https://periodic.com.

[16] URL: https://axiommath.ai.

[17] URL: https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/.

[18] URL: https://www.pewresearch.org/short-reads/2025/06/25/34-of-us-adults-have-used-chatgpt-about-double-the-share-in-2023/.

[19] URL: https://x.com/KobeissiLetter/status/1976276108007092382.

[20] URL: https://www.inquirer.com/politics/pennsylvania/three-mile-island-power-plant-trump-20251120.html.

[21] URL: https://www.whitehouse.gov/fact-sheets/2025/11/fact-sheet-president-donald-j-trump-solidifies-economic-and-defense-partnership-with-the-kingdom-of-saudi-arabia/.

[22] URL: https://www.latitudemedia.com/news/trumps-westinghouse-nuclear-deal-comes-with-unresolved-questions/.