# The AI Force Awakens

**Uli Hitzel**
Executive Geek

April 27, 2024

*Remember how OpenAI's GPT was the default choice for most AI projects? Now, the AI landscape has been flipped on its head in just a few months, with tech giants like Meta, Microsoft, Databricks, and Apple releasing their own open models. Let's look at how this new era of choice and accessibility is changing the way we build smart applications, why there is no one-size-fits-all, and how you can navigate this rapidly evolving landscape to find the best approach that fits your needs.*

Until end of last year, when asked about building smart applications using AI, I would have reluctantly admitted that

OpenAI's GPT from was probably the only viable option at the time. It was like having a toolbox with just one tool – or let's say many powerful tools, but all are from the same brand, with the same limitations and quirks. I'm deeply familiar with the GPT family, having worked with GPT-2, GPT-3, and all the .5 and turbo variants. Make no mistake, they're great, but like I said, it's all one company. I've previously written about the need for a plan B. What happens when your go-to large language model API is suddenly deprecated and replaced with a new one?

## AI Model Revolution: A Diverse Ecosystem Emerges

But then, in just a few months, everything changed. Anthropic and Google have released advanced AI models that are nearly as capable as OpenAI's. Even more significantly, tech giants like Meta, Apple, and Microsoft have made their models open source and available for download. This means you can run them on your own infrastructure instead of just accessing them via API, which some folks aren't comfortable with since it involves sending your data off into the unknown.

Suddenly, my toolbox was overflowing with shiny new tools, each with its own unique features and capabilities, and the AI model landscape had transformed into a diverse ecosystem. Some of these new open models, like Llama 3 are now as powerful as GPT was just 1.5 years ago, and some, incredibly, could even run on my MacBook Air! It was

like watching a black-and-white world suddenly burst into color as AI became more accessible and versatile than ever before. Let's take a look at some of the most important releases from the last few months.

## Recent Advancements in Large Language Models

- **Google Gemini** (December 6, 2023): Google's advanced model lineup, including versions 1.0, 1.5 (Pro and Ultra) with context windows of up to 1 million tokens – I have worked with these quite a bit, and really like how they're developing.
- **Mixtral** from Mistral (December 11, 2023): releasing Mixtral 8x7B, a high-quality sparse mixture of experts model (SMoE) with open weights that matches or outperforms GPT3.5 on most standard benchmarks. Until that time I hadn't explored the Mistral models much yet.
- **Gemma by Google** (February 21, 2024): A family of lightweight, state-of-the-art open models built from the same research and technology as Google's Gemini models, making advanced AI more accessible to developers. They're remarkably small models, and when I tried them out, I realized that for some use cases, we really just need an engine that can understand language and engage with users. Now, we can have those capabilities baked into software that I

can run on my own machine, and I don't need to rely on OpenAI.

- **Anthropic Claude 3** (March 4, 2024): Anthropic's latest offerings include Haiku, Sonnet, and Opus, and I've been fortunate to have research access to these for some time now. In my opinion, they're the most powerful alternatives to OpenAI's models at the moment. Especially Opus feels warm, engaging, and is an incredible model for building digital personalities that can engage in deep intellectual conversations and display advanced reasoning capabilities.
- **Grok by X** (released March 18, 2024): Announced as an "Open Source language model", the company has only released weights to the public, without code or much documentation, leaving many developers uncertain about its actual capabilities and usability. I would be keen to try it, but it's way too big to run on my own hardware or on the cloud at the moment.
- **MM1 by Apple** (released March 19, 2024): a multimodal large language model that clearly indicates the fact that Apple is taking AI very seriously. I have not yet have the chance to try it out.
- **DBRX by Databricks** (March 27, 2024): An open model that introduces a fine-grained mixture-of-experts (MoE) architecture, achieving impressive performance with fewer parameters than GPT. This could be very interesting for enterprise companies, as it gives them the ability to run powerful AI models on their own infrastructure. Databricks' focus on enterprise needs

such as data compliance and the ability to fine-tune the models for specific use cases makes DBRX a compelling option for many businesses.

- **Llama 3** by Meta (April 18, 2024): Meta's latest open model, trained on a massive dataset and offering performance comparable to GPT from just 1.5 years ago. I am very, very impressed by the quality of output, how it reacts to custom prompts, and how versatile and usable it is. Of course I had seen LLama 2 before but did not use it much. Facebook, WhatsApp, and Instagram now have LLama 3 powered capabilities built in. When you think about how getting fresh data to train their models on is a huge problem for AI companies right now, consider how many billions of users these platforms have. Meta is in a great position. Reinforcement learning with human feedback is the key to making AI models smarter and creating the next generation.
- **Phi-3** by Microsoft (April 23, 2024): A family of small language models (SLMs) designed for efficiency and performance, particularly well-suited for edge computing and offline scenarios. I enjoyed my first set of interactions with the Phi-3 models and I believe they will find a strong niche in applications that prioritize speed, privacy, and on-device capabilities over raw power.

This timeline shows the rapid pace of innovation and the increasing diversity of the AI model landscape over just a few months. Each release brings new capabilities,

architectures, and potential use cases, giving developers and businesses a wealth of options to choose from.

## OpenAI and the Rise of More Efficient AI Architectures

But where is OpenAI in all this? The company that runs the ultra prominent ChatGPT and still dominates most of the AI landscape with its GPT models has been notably absent from the recent flurry of releases (if you ignore the SORA announcement for a moment). While they may be working on GPT-5 behind the scenes, it's clear that some of the newer architectures, particularly those using mixture-of-experts (MoE) techniques, are beginning to challenge GPT's supremacy.

Models like DBRX and Llama 3 have demonstrated that it's possible to achieve impressive performance with fewer parameters and more efficient architectures. And the shift towards MoE models, which can deliver better results with smaller infrastructure requirements, is effectively eating OpenAI's lunch by eroding the competitive advantage that GPT once held.

So – how do we navigate this new era of AI model choice and accessibility to find the best fit for our needs?

## Leveraging the Diversity of AI Models

As we navigate this new era of AI model choice and accessibility, it's essential to recognize that there is no one-

size-fits-all solution. Just as our brains employ different systems for various tasks, ranging from the automatic and unconscious (like tying our shoes) to the deliberate and analytical (like solving a complex math problem), we can leverage different AI models for specific purposes.

In his book "Thinking, Fast and Slow," psychologist Daniel Kahneman introduces the concept of two distinct systems in our minds: System 1, which operates automatically and quickly, and System 2, which allocates attention to more effortful mental activities. We can apply this framework to the world of AI models, using smaller, more efficient models for tasks that require quick responses and minimal computational resources (akin to System 1) and larger, more complex models for tasks that demand deeper reasoning and analysis (akin to System 2).

Moreover, we can combine models that excel at specific tasks to create powerful, multi-faceted AI systems. For example, we might use a model like Phi-3 for edge computing and offline scenarios, Gemma for lightweight, on-device processing, and DBRX or Llama 3 for more complex, cloud-based tasks. By understanding the strengths and weaknesses of each model and strategically combining them, we can build AI applications that are more efficient, effective, and adaptable to a wide range of use cases.

## You Can Use A Flexible, Informed Approach

Ultimately, the key to success in this new landscape is to approach AI model selection and deployment with a flexible, informed mindset. By staying up-to-date on the latest developments, experimenting with different models, and carefully considering the specific needs and constraints of each project, developers and businesses can harness the power of this new era of AI while avoiding the pitfalls of over-reliance on any single model or provider.

It's an exciting time for AI, and as we move forward, let us embrace the diversity and accessibility of the AI model landscape, as this clearly represents a significant step forward in the democratization and advancement of artificial intelligence.