The Next Token of Progress: 4 Unlocks on the Generative AI Horizon

by Sarah Wang and Shangda Xu

Large language models (LLMs) have taken the tech industry by storm, powering experiences that can only be described as magical—from writing a week's worth of code in seconds to generating conversations that feel even more empathetic than the ones we have with humans. Trained on trillions of tokens of data with clusters of thousands of GPUs, LLMs demonstrate remarkable natural language understanding and have transformed fields like copy and code, propelling us into the new and exciting generative era of AI. As with any emerging technology, generative AI has been met with some criticism. Though some of this criticism does reflect current limits of LLMs' current capabilities, we see these roadblocks not as fundamental flaws in the technology, but as opportunities for further innovation.

To better understand the near-term technological breakthroughs for LLMs and prepare founders and operators for what's around the bend, we spoke to some of the leading generative AI researchers who are actively building and training some of the largest and most cutting edge models: Dario Amodei, CEO of Anthropic; Aidan Gomez, CEO of Cohere; Noam Shazeer, CEO of Character.AI; and Yoav Shoham of AI21 Labs. These conversations identified 4 key innovations on the horizon: **steering, memory, "arms and legs,"** and **multimodality.** In this piece, we discuss how these key innovations will evolve over the next 6 to 12 months and how founders curious about integrating AI into their own businesses might leverage these new advances.

Steering

Many founders are understandably wary of implementing LLMs in their products and workflows because of these models' potential to hallucinate and reproduce bias. To address these concerns, several of the leading model companies are working on improved **steering**—a way to place better controls on LLM outputs—to focus model outputs and help models better understand and execute on complex user demands. Noam Shazeer draws a parallel between LLMs and children in this regard: "it's a question of how to direct [the model] better… We have this problem with LLMs that we just need the right ways of telling them to do what we want. Small children are like this as well—they make things up sometimes and don't have a firm grasp of fantasy versus reality." Though there has been notable progress in steerability among the model providers as well as the emergence of tools like Guardrails and

LMQL, researchers are continuing to make advancements, which we believe is key to better productizing LLMs among end users.

Improved steering becomes especially important in **enterprise** companies where the consequences of unpredictable behavior can be costly. Amodei notes that the unpredictability of LLMs “freaks people out” and, as an API provider, he wants to be able to “look a customer in the eye and say ‘no, the model will not do this,’ or at least does it rarely.” By refining LLM outputs, founders can have greater confidence that the model’s performance will align with customer demands. Improved steering will also pave the way for broader adoption in other industries with higher accuracy and reliability requirements, like advertising, where the stakes of ad placement are high. Amodei also sees use cases ranging from “legal use cases, medical use cases, storing financial information and managing financial bets, [to] where you need to preserve the company brand. You don’t want the tech you incorporate to be unpredictable or hard to predict or characterize.” With better steering, LLMs will also be able to do more complex tasks with less prompt engineering, as they will be able to better understand overall intent.

Advances in LLM steering also have the potential to unlock new possibilities in sensitive **consumer** applications where users expect tailored and accurate responses. While users might be willing to tolerate less accurate outputs from LLMs when engaging with them for conversational or creative purposes, users want more accurate outputs when using LLMs to assist them in daily tasks, advise them on major decisions, or augment professionals like life coaches, therapists, and doctors. Some have pointed out that LLMs are poised to unseat entrenched consumer applications like search, but we likely need better steering to improve model outputs and build user trust before this becomes a real possibility.

Key unlock: users can better tailor the outputs of LLMs.

Memory

Copywriting and ad-generating apps powered by LLMs have already seen great results, leading to quick uptake among marketers, advertisers, and scrappy entrepreneurs. Currently, however, most LLM outputs are relatively generalized, which makes it difficult to leverage them for use cases requiring personalization and contextual understanding. While prompt engineering and fine-tuning can offer some level of personalization, prompt engineering is less scalable and fine-tuning tends to be expensive, since it requires some degree of re-training and often partnering closely with mostly closed source LLMs. It’s often not feasible or desirable to fine-tune a model for every individual user.

In-context learning, where the LLM draws from the content your company has produced, your company's specific jargon, and your specific context, is the holy grail—creating outputs that are more refined and tailored to your particular use case. In order to unlock this, LLMs need enhanced memory capabilities. There are two primary components to LLM memory: **context windows** and **retrieval**. Context windows are the text that the model can process and use to inform its outputs *in addition to* the data corpus it was trained on. Retrieval refers to retrieving and referencing relevant information and documents from a body of data outside the model's training data corpus ("contextual data"). Currently, most LLMs have limited context windows and aren't able to natively retrieve additional information, and so generate less personalized outputs. With bigger context windows and improved retrieval, however, LLMs can directly offer much more refined outputs tailored to individual use cases.

With expanded **context windows** in particular, models will be able to process larger amounts of text and better maintain context, including maintaining continuity through a conversation. This will, in turn, significantly enhance models' ability to carry out tasks that require a deeper understanding of longer inputs, such as summarizing lengthy articles or generating coherent and contextually accurate responses in extended conversations. We're already seeing significant improvement with context windows—GPT-4 has both an 8k and 32k token context window, up from 4k and 16k token context windows with GPT-3.5 and ChatGPT, and Claude recently expanded its context window to an astounding 100k tokens.

Expanded context windows alone don't sufficiently improve memory, since cost and time of inference scale quasi-linearly, or even quadratically, with the length of the prompt. **Retrieval** mechanisms augment and refine the LLM's original training corpus with contextual data that is most relevant to the prompt. Because LLMs are trained on one body of information and are typically difficult to update, there are two primary benefits of retrieval according to Shoham: "First, it allows you to access information sources you didn't have at training time. Second, it enables you to focus the language model on information you believe is relevant to the task." Vector databases like Pinecone have emerged as the de facto standard for the efficient retrieval of relevant information and serve as the memory layer for LLMs, making it easier for models to search and reference the right data amongst vast amounts of information quickly and accurately.

Together, increased context windows and retrieval will be invaluable for **enterprise** use cases like navigating large knowledge repositories or complex databases. Companies will be able to better leverage their proprietary data, like internal knowledge, historical customer support tickets, or financial results as inputs to LLMs without fine-tuning. Improving LLMs' memory will lead to improved and deeply customized capabilities in areas like training, reporting, internal search, data analysis and business intelligence, and customer support.

In the **consumer** space, improved context windows and retrieval will enable powerful personalization features that can revolutionize user experiences. Noam Shazeer believes that “one of the big unlocks will be developing a model that both has a very high memory capacity to customize for each user but can still be served cost-effectively at scale. You want your therapist to know everything about your life; you want your teacher to understand what you know already; you want a life coach who can advise you about things that are going on. They all need context.” Aidan Gomez is similarly excited by this development. “By giving the model access to data that’s unique to you, like your emails, calendar, or direct messages,” he says, “the model will know your relationships with different people and how you like to talk to your friends or your colleagues and can help you within that context to be maximally useful.”

Key unlock: LLMs will be able to take into account vast amounts of relevant information and offer more personalized, tailored, and useful outputs.

“Arms and legs”: giving models the ability to use tools

The real power of LLMs lies in enabling natural language to become the conduit for action. LLMs have a sophisticated understanding of common and well-documented systems, but they can’t execute on any information they extract from those systems. For example, OpenAI’s ChatGPT, Anthropic’s Claude, and Character AI’s Lily can describe, in detail, how to book a flight, but they can’t natively book that flight themselves (though advancements like ChatGPT’s plugins are starting to push this boundary). “There’s a brain that has all this knowledge in theory and is just missing the mapping from names to the button you press,” says Amodei. “It doesn’t take a lot of training to hook those cables together. You have a disembodied brain that knows how to move, but it doesn’t have arms or legs attached yet.”

We’ve seen companies steadily improve LLMs’ ability to use tools over time. Incumbents like Bing and Google and startups like Perplexity and You.com introduced search APIs. AI21 Labs introduced Jurassic-X, which addressed many of the flaws of standalone LLMs by combining models with a predetermined set of tools, including a calculator, weather API, wiki API, and database. OpenAI betaed plugins that allow ChatGPT to interact with tools like Expedia, OpenTable, Wolfram, Instacart, Speak, a web browser, and a code interpreter—an unlock that drew comparisons to Apple’s “App Store” moment. And more recently, OpenAI introduced function calling in GPT-3.5 and GPT-4, which allows developers to link GPT’s capabilities to whatever external tools they want.

By shifting the paradigm from knowledge excavation to an action orientation, adding arms and legs has the potential to unlock a range of use cases across companies and user types. For **consumers**, LLMs may soon be able to give you recipe ideas then order the groceries you need, or suggest a brunch spot and book your table. In the **enterprise**, founders can make their apps easier to use by plugging in LLMs. As Amodei notes, “for features that are very hard to use from a UI perspective, we may be able to make complicated things happen by just describing them in natural language.” For instance, for apps like Salesforce, LLM integration should allow users to give an update in natural language and have the model automatically make those changes—significantly cutting down the time required to maintain the CRM. Startups like Cohere and Adept are working on integrations into these kinds of complex tools.

Gomez believes that, while it’s increasingly likely that LLMs will be able to use apps like Excel within 2 years, “there’s a bunch of refinement that still needs to happen. We’ll have a first generation of models that can use tools that will be compelling but brittle. Eventually, we’ll get the dream system, where we can give any software to the model with some description of ‘here’s what the tool does, here’s how you use it’, and it’ll be able to use it. Once we can augment LLMs with specific and general tools, the sort of automation it unlocks is the crown jewel of our field.”

Key unlock: LLMs will be able to interact much more effectively with the tools we use today.

Multimodality

While the chat interface is exciting and intuitive for many users, humans hear and speak language as or more often than they write or read it. As Amodei notes, “there is a limit to what AI systems can do because not everything is text.” Models featuring **multimodality**, or the ability to seamlessly process and generate content across multiple audio or visual formats, changes this interaction to beyond language. Models like GPT-4, Character.AI, and Meta’s ImageBind already process and generate images, audio, and other modalities, but they do so at a more basic—though quickly improving—level. In Gomez’s words, “our models are blind in a literal sense today—that needs to change. We’ve built a lot of graphical user interfaces (GUIs) that assume [the user] can see.”

As LLMs evolve to better understand and interact with multiple modalities, they’ll be able to use existing apps that rely on GUIs today, like the browser. They can also offer more engaging, connected, and comprehensive experiences to consumers, who will be able to engage outside of a chat interface. “A lot of great integration with multimodal models can make things a lot more engaging and connected to the user,” Shazeer points out. “I believe, for now, most of the core intelligence comes from text, but audio and video can make these things more fun.” From video chats with AI tutors to iterating and writing TV pilot scripts with an AI partner, multimodality has the potential to change entertainment, learning and development, and content generation across a variety of consumer and enterprise use cases.

Multimodality is also closely tied to tool use. While LLMs might initially connect with outside software through APIs, multimodality will enable LLMs to use tools designed for humans that don't have custom integrations, like legacy ERPs, desktop applications, medical equipment, or manufacturing machinery. We're already seeing exciting developments on this front: Google's Med-PaLM-2 model, for instance, can synthesize mammograms and X-rays. And as we think longer-term, multimodality—particularly integration with computer vision—can extend LLMs into our own physical reality through robotics, autonomous vehicles, and other applications that require real-time interaction with the physical world.

Key unlock: Multimodal models can reason about images, video, or even physical environments without significant tailoring.

While there are real limitations to LLMs, researchers have made astounding improvements to these models in a short amount of time—in fact, we've had to update this article multiple times since we started writing it, a testament to the lightning-fast progression of this technology in the field. Gomez agrees: "An LLM making up facts 1 in 20 times is obviously still too high. But I really still feel quite confident that it's because this is the first time we've built a system like that. People's expectations are quite high, so the goal post has moved from 'computer is dumb and does only math' to 'a human could've done this better'. We've sufficiently closed the gap so that criticism is around what a human can do."

We're particularly excited about these 4 innovations, which are on the cusp of changing the way founders build products and run their companies. The potential is even greater in the long term. Amodei predicts that, "at some point, we could have a model that will read through all the biological data and say: here's the cure for cancer." Realistically, the best new applications are likely still unknown. At Character.AI, Shazeer lets the users develop those use cases: "We'll see a lot of new applications unlocked. It's hard for me to say what the applications are. There will be millions of them and the users are better at figuring out what to do with the technology than a few engineers." We can't wait for the transformative effect these advancements will have on the way we live and work as founders and companies are empowered with these new tools and capabilities.