DOI:10.1145/3615859

Michael A. Cusumano

Technology Strategy and Management

Generative AI as a New Innovation Platform

Considering the stability and longevity of a potential new foundational technology.

RISING ATTENTION ABOUT generative AI prompts the question: Are we witnessing the birth of a new innovation platform? The answer seems to be yes, though it remains to be seen how pervasive this new technology will become.

To have an innovation platform, there must be a foundational technology, such as a widely adopted personal computer or smartphone operating system, or the Internet and cloud-computing services with application programming interfaces (APIs) (see "The Cloud as an Innovation Platform for Software Development," *Communications*, October 2019). Third parties are then needed to access these APIs and start creating complementary products and services. More applications attract more users, which leads to more applications and then more users, and usually improvements in the foundational technology. Self-reinforcing positive feedback loops ("network effects") between users and applications usually indicate the presence of a new innovation platform and supporting ecosystem.[12,26]

Generative AI itself is not a complete platform but rather a powerful enabling technology, based on a specific type of neural network and machine learning. It has taken decades to develop, but progress greatly accelerated and changed direction due to innovative research done at Google and published in 2017.[41] A team of scientists designed a neural network that could identify patterns in language (rather than analyzing words one by one) and transform the analysis into predictions of what words or phrases should come next. There are many potential applications for such a technology beyond text translation. The key researchers moved on to several firms, including OpenAI, creator of ChatGPT (GPT stands for "generative pre-trained transformer").[33]

Bloomberg estimated the market for generative AI hardware and software was already worth $40 billion in 2022 and likely to grow to $1.3 trillion over the next 10 years.[2] ChatGPT alone may have attracted as many as one billion users as of July 2023.[8] Usage levels seem to be slowing.[14] But at least 335 startups now target generative AI.[7] Established companies are exploring ways to incorporate generative AI into existing products and services. Who are the key players and how are they organized? What are the opportunities and what should concern us?

Structure of the Generative AI Ecosystem

OpenAI was established in 2015 and now benefits from $10 billion in funding from Microsoft. It "productized" and then "platformized" generative AI technology when it introduced GPT-3 in 2020, ChatGPT in 2022, and then GPT-4 in 2023, with accessible "chatbots" (conversational interfaces—the product) as well as APIs (developer interfaces—the platform).[17] This whole class of AI systems "generate" text, graphics,

IMAGE BY LOGIN/SHUTTERSTOCK

audio, and video using learning algorithms based on large language models (LLMs) that train on huge datasets (almost whole "languages").[29] The chatbot responses to queries are humanlike, but supercharged with enormous computer processing power and access to trillions of words and other data points.

The rapidly growing ecosystem has several layers: Joining OpenAI are several other producers of foundational models, which are similar to operating systems married to neural networks and analytics, with APIs. Then we have infrastructure providers, which offer specialized hardware and cloud-computing services; and applications developers, both "horizontal" (targeting a broad set of users) and "vertical" (targeting specific industries).

▸ Foundational models: Users can access generative AI chatbots through Internet browsers, but the underlying development environment is the LLM software. There is a lot of competition (unlike in PC or smartphone operating systems) because we are at an early stage. OpenAI and Microsoft (ChatGPT, Bing), Google (DeepMind, Bard, AlphaFold), and Meta (LLaMA 2) have attracted the most users and developers.[19] Other big firms include Amazon (Alexa, AWS), Alibaba (DAMO), and Baidu (Ernie Bot). Then we have well-funded startups such as Cohere, A121 Labs, Contextual AI, Hugging Face, Anthropic, and Inflection AI.[22] Network effects around application compatibility tend to drive digital platform markets toward a small number of big winners.[12,16] This will likely happen with generative AI, though the competing LLMs are now trying to differentiate themselves or find a niche to survive. Some target consumer versus enterprise users, or broad versus focused application support.

▸ Infrastructure: Most PC, smartphone, and Internet software runs on central processing units (CPUs) designed by Intel, ARM, or AMD. However, generative AI systems run best on graphical processing units (GPUs) optimized for massively parallel processing, such as for gaming or complex mathematics. The leader in GPUs for generative AI, with approximately 80% of the market, is Nvidia, which recently passed $1 trillion in market value.[4] Nvidia also offers software development tools and frameworks to help application developers use (and keep using) its hardware.[29] The cloud-computing providers, led by Amazon, Microsoft, and Google, all must support generative AI systems with Nvidia GPUs in their datacenters. These cloud vendors offer their own LLMs via APIs as well. They are also looking at an enormous new revenue stream since all the LLMs use massive computing resources.[22] In parallel, demand and prices for GPU chips have skyrocketed and stimulated competition. Intel bought Habana Labs for $2 billion in 2019 to improve its GPU offerings. Amazon, Microsoft, and Google are designing custom GPUs for their datacenters. Several

GPU startups have attracted billions of dollars in funding.[3]

▸ Horizontal applications: Microsoft and Google have already added LLMs to their search engines, enabling billions of people to access this technology with ease.[21,28] In May 2023, Microsoft also released plug-ins to connect OpenAI technology embedded in Microsoft 365 Copilot with business applications from various vendors.[42] Other firms are doing the same. Plug-ins enable generative AI systems to access customer data and write reports or trigger actions in other programs. Meanwhile, many startups are introducing tools for text, image, audio, video, and code generation, as well as chatbot design, search, data management, and machine learning.[19,22]

▸ Vertical applications: Generative AI startups are already building specific applications for a growing variety of industries. These include manufacturing, gaming, fashion, retail, energy, healthcare, defense, finance, agriculture, physical infrastructure, education, media and entertainment, legal services, computer coding, mobility, and construction.[6,22]

What to Worry About: Regulation and Governance

Most dominant platform companies have provoked antitrust scrutiny, with mixed results. We also have seen company and user behavior that is difficult to control and has engendered broad mistrust in digital platforms and content ("Section 230 and a Tragedy of the Commons: The Dilemma of Social Media Platforms," *Communications*, October 2021).

The challenges of generative AI are similar to what we have seen before but potentially more difficult to resolve. Geoffrey Hinton, a pioneer in machine learning for neural networks, left his position at Google in May 2023 after warning generative AI would diffuse too much misinformation and become detrimental to society. He especially worried that the current systems had no guardrails limiting bad behavior or social damage.[31] There are several related concerns that must be addressed:

Concentration of market power. Two forces are at play here. First, we are likely to see a reduction in the number of competing LLMs as developers choose the most popular or accessible models around which to build their applications. Second, only a small number of companies have the money to keep developing the foundational models *and* fund the enormous computing resources required to offer generative AI as a cloud service. The partially open source LLM from Meta (Facebook) provides an alternative, but it still requires funding and a cloud partner (currently Microsoft Azure).[10,25] It seems unlikely an open source platform or small players will be able to compete long-term with giant firms such as Microsoft and Google without some governmental or industry-level interventions.

Content ownership and privacy. We have encountered data privacy, bias, and content ownership issues with prior digital platforms. For Internet search, a U.S. appeals court ruled in 2008 that a few lines of text—but not more—was a "fair use" of copyrighted content.[18] Generative AI takes the use of other people's data and images to another level. It is already a matter of litigation that LLM producers are not compensating creators of the content that feeds their learning algorithms.[15,39] There is a lawsuit challenging how Microsoft, GitHub, and OpenAI have learned to produce computer code from copyrighted open source software.[38] Italy temporarily banned ChatGPT due to privacy concerns.[34] We know there is bias built into AI algorithms and the data they use to learn.[5,23] Difficult ownership challenges will arise whenever generative AI systems seem to "invent" their own content.[30] We already see teachers struggling with how to deal with homework assignments produced or enhanced by generative AI.[32] Companies and other organizations can address some of these concerns with internal policies. However, courts and governments will have to settle legal disputes and answer the new trillion-dollar question: What is "fair use" of training data for generative AI systems?

Information accuracy and authenticity. When LLMs cannot find an answer to a query, they use predictive analytics to make up reasonable but sometimes incorrect responses, called hallucinations.[13] This problem should lessen with better technology and design poli-

What is "fair use" of training data for generative AI systems?

cies. For example, it is possible to direct LLMs to check their sources or to use only particular content.[9] However, human beings themselves dispute interpretations of the same facts and data. Generative AI systems may not be any better, particularly if they base analyses on false or ambiguous information. This is also a business opportunity: Various startups offer tools to help users detect fake text, audio, and video.[24] Yet, so far, it does not seem these tools can reliably distinguish genuine from false text (or any other digital content).[37] Meanwhile, generative AI systems keep improving—maybe exponentially.

Regulation versus self-regulation. Some industries, such as movies and video games, advertising on television and radio, and airline reservations, have effectively combined government regulation, or the credible threat of regulation, with company efforts to regulate themselves.[11] Generative AI systems will need a similar combination of oversight and self-regulation. The U.S. Federal Trade Commission already has opened an investigation into ChatGPT's inaccurate claims and data leaks, though it is unclear what laws apply.[43] In July 2023, the White House announced that Google, Amazon, Microsoft, Meta, OpenAI, Anthropic, and Inflection AI all agreed to allow independent security testing of their systems and data as well as to add digital watermarks to generative AI images, text, and video.[44] Adobe heads the Content Authenticity Initiative, a consortium of 1,000 companies and other organizations that is trying to establish standards to help detect fake content.[24] These are all positive steps. Nonetheless, company promises to regulate themselves are usually insufficient, especially with new, rapidly evolving technologies.[27] Open source platforms could help but they are double-edged swords: More competitors and "eyeballs" may reduce big-firm dominance and help expose technical or policy flaws. But bad actors will also have access to open source technology.

Environmental impact. Some new platforms, such as Bitcoin and blockchain, consume enormous amounts of energy. Generative AI is likely to take energy consumption to another level. Chatbots have mass-market appeal and there is almost unlimited potential for applications. Computing resources required for LLM training and then responses to each chatbot prompt are already huge. By some estimates, generative AI's use of computing resources has been increasing exponentially for years, doubling every 6 to 10 months.[36,40]

Unintended consequences. No one knows where this new technology will lead us. At the least, many occupations (teachers, journalists, lawyers, travel agents, stock traders, actors, computer programmers, corporate planners … military strategists?) may find their jobs replaced, enhanced, or greatly altered.

Generative AI may turn out to be less important or disruptive than it seems at present.[1] Still, as Thomas Friedman wrote in *The New York Times*, this is "our Promethean moment."[20] Now is the time to shape the future of this new platform and ecosystem, before the technology becomes more deeply entrenched in our personal and professional lives. C

References

1. Bishop, T. Bill Gates: AI breakthroughs are the biggest tech advance since the graphical user interface. *Geekwire* (Mar. 21, 2023).
2. Bloomberg Intelligence. Generative AI to become a $1.3 trillion market by 2032, research finds. *Bloomberg.com* (June 1, 2023).
3. Bradshaw, T. Startups seek to challenge Nvidia's dominance over AI chip market. *Financial Times* (July 21, 2023).
4. Bradshaw, T. and Waters, R. How Nvidia created the chip powering the generative AI boom. *Financial Times* (May 26, 2023).
5. Buell, S. An MIT student asked AI to make her headshot more 'professional.' It gave her lighter skin and blue eyes. *Boston Globe* (July 19, 2023).
6. CB Insights. AI 100: The most promising artificial intelligence startups of 2023. CBInsights.com (June 20, 2023).
7. CB Insights. The generative AI market map: 335 vendors automating content, code, design, and more. CBInsights.com (July 12, 2023).
8. CB Insights. The state of LLM developers in 6 charts. CBInsights.com (July 14, 2023).
9. Chen, B. We're using A.I. chatbots wrong. Here's how to direct them. *New York Times* (July 20, 2023).
10. Criddle, C. et al. Meta to release commercial AI model in effort to catch rivals. *Financial Times* (July 13, 2023).
11. Cusumano, M. et al. Can self-regulation save digital platforms? *Industrial and Corporate Change 20*, 5 (Oct. 2021), 1259–1285.
12. Cusumano, M. et al. *The Business of Platforms.* Harper Business, New York, 2019.
13. De Vynck, G. ChatGPT 'hallucinates.' Some researchers worry it isn't fixable. *Washington Post* (May 30, 2023).
14. De Vynck, G. Every start-up is an AI company now. Bubble fears are growing. *Washington Post* (Aug. 5, 2023).
15. De Vynck, G. AI learned from their work. Now they want compensation. *Washington Post*, (July 16, 2023).
16. Eisenmann, T. et al. Strategies for two-sided markets. *Harvard Business Review 84*, 10 (Oct. 2006), 92–101.
17. Enterprise DNA Experts. What is the ChatGPT API: An essential guide. Blog.enterprisedna.co (July 19, 2023).
18. EveryCRSReport. Internet search engines: Copyright's 'fair use' in reproduction and display rights. EveryCRSReport.com (Jan. 9, 2007–Jan. 28, 2008).
19. Forsyth, O. Mapping the generative AI landscape. Antler.com (Dec. 20, 2022).
20. Friedman, T. Our new Promethean moment. *The New York Times* (Mar. 21, 2023).
21. Grant, N. Google devising radical search changes to beat back A.I. rivals. *The New York Times* (Apr. 16, 2023).
22. Greenman, S. Who will make money from the generative AI gold rush? Part I. Medium.com (Mar. 12, 2023).
23. Hill, K. OpenAI worries about what its chatbot will say about people's faces. *The New York Times* (July 18, 2023).
24. Hsu, T. and Myers, S. Another side of the A.I. boom: Detecting what A.I. makes. *The New York Times* (May 18, 2023).
25. Isaac, M. and Metz, C. Meta unveils a more powerful A.I. and isn't fretting over who uses it. *The New York Times* (July 18, 2023).
26. Jacobides, M. et al. Towards a theory of ecosystems. *Strategic Management J. 39*, 8 (May 2018), 2255–2276.
27. Kang, C. In U.S., regulating A.I. is in its 'early days.' *New York Times* (July 21, 2023).
28. Kruppa, M. Google plans to make search more 'personal' with AI chat and video clips. *Wall Street Journal* (May 6, 2023).
29. Lee, A. What are large language models used for?" Nvidia Blog (Jan. 26, 2023).
30. Lohr, S. Can A.I. invent? *New York Times* (July 15, 2023).
31. Metz, C. The 'godfather of A.I.' leaves Google and warns of danger ahead. *New York Times* (May 1, 2023).
32. Mollick, E. The homework apocalypse. oneusefulthing.org (July 1, 2023).
33. Murgia, M. Transformers: The Google scientists who pioneered a revolution. *Financial Times* (July 23, 2023).
34. Murgia, M. and Sciorilli Borrelli, S. Italy temporarily bans ChatGPT over privacy concerns. *Financial Times* (Mar. 31, 2023).
35. Noble, C. Generative AI's hidden cost: Its impact on the environment. Nasdaq.com (June 20, 2023).
36. Oremus, W. AI chatbots lose money every time you use them. That is a problem. *Washington Post* (June 5, 2023).
37. Sadasivan, V. et al. Can AI-generated text be reliably detected? Department of Computer Science, University of Maryland. arxiv.org (June 28, 2023).
38. Samuelson, P. Legal challenges to generative AI, part I. *Commun. ACM 66*, 7 (July 2023).
39. Small, Z. Sarah Silverman sues OpenAI and Meta over copyright infringement. *New York Times* (July 10, 2023).
40. The bigger-is-better approach to AI is running out of road. *The Economist* (June 21, 2023).
41. Vaswani, A. et al. Attention is all you need. In *Proceedings of the 31st Conf. on Neural Information Processing Systems* (NIPS 2017).
42. Waters, R. Microsoft launches generative AI tools for developers. *Financial Times* (May 24, 2023).
43. Zakrzewski, C. FTC investigated OpenAI over data leak and ChatGPT's inaccuracy. *Washington Post* (July 13, 2023).
44. Zakrzewski, C. Top tech firms sign White House pledge to identify AI-generated images. *Washington Post* (July 21, 2023).

Michael A. Cusumano (cusumano@mit.edu) is a professor and Deputy Dean at the Massachusetts Institute of Technology Sloan School of Management, Cambridge, MA, USA, coauthor of *The Business of Platforms* (2019), and a member of the MIT Center for Quantum Engineering (https://cqe.mit.edu/).

The author thanks Annabelle Gawer and David Yoffie for their comments.