THE UNIVERSITY OF MELBOURNE

Centre for Artificial Intelligence and Digital Ethics

CAIDE AI Policy Forums

FORUM #3 ISSUES PAPER: Competition and Markets

About the Author

Calvin Collins is a Research Assistant at the Centre for Artificial Intelligence and Digital Ethics (CAIDE) at the University of Melbourne. Calvin completed his Bachelor of Arts (2017) and his Juris Doctor (2021) at Melbourne Law School, before becoming a paralegal, and eventually a solicitor, at leading global law firm Herbert Smith Freehills. At CAIDE, Calvin pairs his legal knowledge and experience with an interest in emerging technology regulation. Calvin would like to acknowledge the contributions of Professor Jeannie Marie Paterson to the editing of this paper.

What is CAIDE?

The Centre for Artificial Intelligence and Digital Ethics (CAIDE) is a cross-disciplinary research centre at the University of Melbourne. CAIDE facilitates cross-disciplinary research, teaching and leadership on the ethical, technical, regulatory and legal issues relating to AI and digital technologies. CAIDE is directed by Professor Jeannie Marie Paterson from Melbourne Law School. For more information about CAIDE, see our website: https://www.unimelb.edu.au/caide.

Acknowledgements and Sponsorships

The AI Policy forums are supported by:

- The Centre for AI and Digital Ethics, funded by the Faculty of Engineering and Information Technology and Melbourne Law School at the University of Melbourne
- The Ninian Stephen Law Program powered by the Menzies Foundation, as part of the project *New Legal Thinking for Emerging Technologies*
- Microsoft, Atlassian and the Tech Council of Australia as part of the project *Demystifying Generative AI*.

CAIDE AI Policy Forums

This issues paper was prepared as reading for the CAIDE AI Policy Forum #3, held at the University of Melbourne in 2024. It should be read in conjunction with the discussion paper on GenAI and Markets, available here. The aim of the AI Policy Forums is to provide an opportunity for discussing policy and law issues raised by the emergence of AI that go beyond the headlines. This event considered the impact of AI, particularly Generative AI, on markets and competition law.

Contact CAIDE

Email: uom-caide@unimelb.edu.au.
Website: https://www.unimelb.edu.au/caide.
LinkedIn: https://au.linkedin.com/company/caide-unimelb.

Demystifying Generative AI and Markets

Generative AI- that is, AI models that can generate language, video or audio from user prompts[1]- has raised excitement and concern across the globe. Included in this concern are questions about the potential impact of generative AI on competition across its supply chain and into related markets. Competition regulators in many jurisdictions have issued statements of intent to investigate these concerns, including the US Federal Trade Commission,[2] the UK Competition and Markets Authority,[3] and the ACCC's latest Digital Platforms Inquiry issues paper.[4]

The generative AI market

The frenetic pace of change in generative AI technologies is reflected in the competitive landscape. As observed by the UK Competition and Markets Authority, over 120 foundation models (multipurpose models trained on broad data sets that can serve as the foundation to various applications of AI) were released between September 2023 and March 2024, bringing the total number of known foundation models globally to over 330.[5]

Although the pace of change means distinctions between different types of models can also shift, there are two key ways by which generative AI models can be differentiated. These include:

1. Large and small language models: Foundation models, such as OpenAI's GPT-4 and Anthropic's Claude, are 'large language models' (**LLMs**). LLMs are trained upon vast swathes of data that equip them with advanced language (and, increasingly, visual and sound) recognition and generation abilities that allow them to undertake tasks across an array of fields, ranging from creative endeavours to scientific research.

   By contrast, small language models (**SLMs**) are trained on comparatively small data sets and use less computation power. SLMs can sometimes be as well-suited to a particular use-case as LLMs, particularly when used delimited subset of tasks that reflect the model's targeted training data.

   Importantly, the boundaries between what is considered a 'large' or 'small' language model are shifting as the technologies develop. For example, when OpenAI's GPT-2 was released, it was considered a large language model with 1.5 billion parameters. Now it is considered small when contrasted against OpenAI's subsequent GPT-4 model, which is estimated to have 1.8 trillion parameters.

2. Open-source and proprietary models: Open-source models are made available on public repositories such as Hugging Face, which can be freely downloaded by developers to use in their applications. Open-source models are being developed by a range of players from start-ups (e.g., Mistral's Mixtral models), large corporations (e.g., Meta's Llama models and Google's Gemma models), and non-profit organizations (e.g., AI2's OLMo models).

   In contrast, closed-source models (also known as 'proprietary models') do not make their underlying weights and biases publicly available to developers.

The rich variety and pace of AI-model development has facilitated competition in the supply of generative AI products and services, as market participants across the AI tech stack vie to produce the most efficient, accurate, user-friendly or specialised technologies. The introduction of generative AI technologies to other products or services has also added a new dimension of competition to the markets for these products.

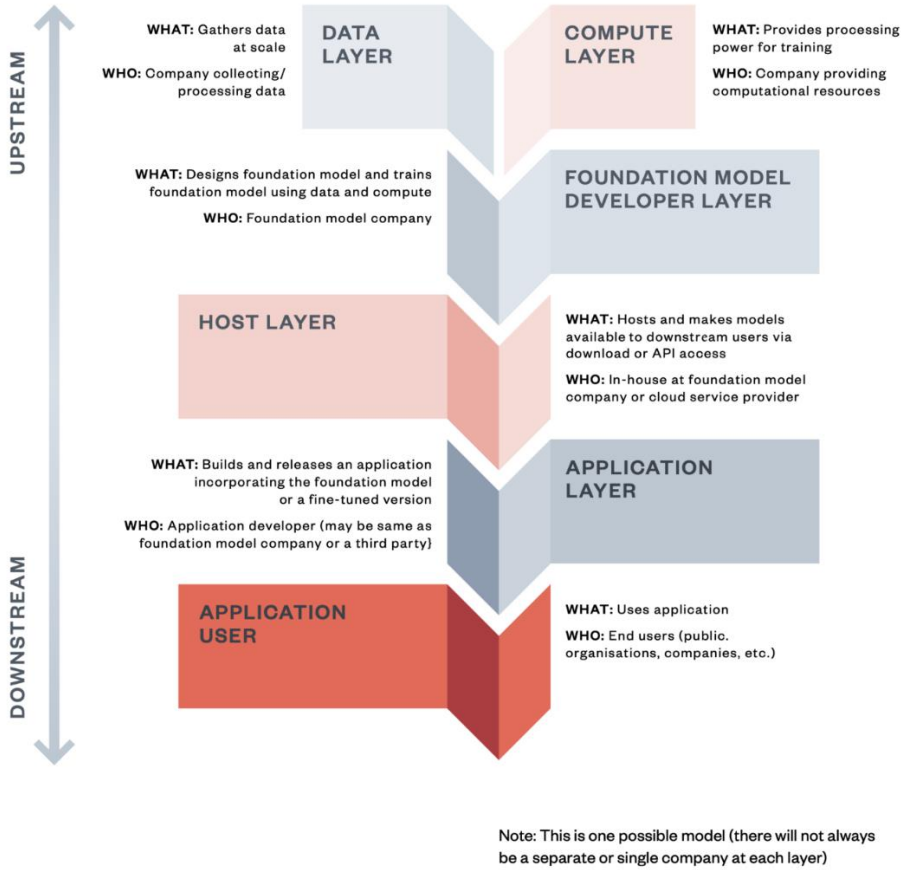
The generative AI supply chain

The diversity of generative AI technologies is also reflected in the variety of AI developers active across each layer of the AI 'stack' (the variety of infrastructure and technologies leveraged to create applications with generative AI capabilities). These include both large, well-funded developers, such as Google, Meta, Apple, NVIDIA, Microsoft, Open AI and IBM, and start-ups like Anthropic, Cohere, Deci, Mistral, Nixtla, Stability AI, TII, Perplexity, Snowflake and xAI.

While the generative AI supply chain can vary significantly across different types of AI models, there are some common features across the tech stack. These include:

1. Datacentre infrastructure, that includes advanced Graphics Processing Units (GPUs) with high bandwidth network connections that enable data gathering and computing power at scale.

2. Leveraging datacentre infrastructure, developers and research scientists to train and develop foundation models (e.g. GPT-4, Midjourney, Claude) for downstream access.
3. Downstream access provided at the host layer via direct download by the model developer or via Application Program Interfaces (APIs) by cloud service providers. APIs provide a way for applications (e.g. ChatGPT, Bing Copilot, Bard) to access and use the output from foundation models to deliver services to individuals and organisations.



Source: Ada Lovelace Institute, Explainer: What is a foundation model? 17 July 20232

Is there currently competition in the GenAI market?

The prevalence of competition varies across the supply chain. At the hardware level, high barriers to entry and geopolitical forces have largely prevented the proliferation of firms capable of producing the chips underpinning generative AI, with Nvidia currently commanding 80% of market share.[6] Similarly, there are significant economies of scale that exist at the datacentre infrastructure and foundation model developer layer, that has seen only a small pool of sophisticated firms dominate the market, such as Meta, Amazon and OpenAI/Microsoft.

However, further down the supply chain at the application layer, increasingly fierce competition has grown as smaller firms compete to develop bespoke and targeted uses for the foundational models. As of August 2024, approximately 370,000 transformer-based models (models utilising a neural network architecture that transforms or changes an input sequence into an output sequence) were listed on the public AI model archive Hugging Face.[7] In the eight months from December 2023 to August 2024, this number increased by 180,000, or 93%, and is continuing to constantly increase.[8]

Many of these smaller players are competing at the application level with larger firms using those same larger firms' foundation models. To date, the dominant market players have condoned this competition in the spirit of promoting innovation in an inchoate industry. For example, in its 'AI Access Principles' published in February 2024, Microsoft committed itself to promoting access to its AI infrastructure, acknowledging that it has "a responsibility to enable innovation and foster competition".[9] However, this democratisation of application development through facilitating access to foundation models at once promotes and bottlenecks competition: firms with a foundation model retain the option to restrict the use of their model and stifle smaller firms' ability to develop and operate competing applications.

Given the multiplicity of products and services offered by these firms, there is also the risk that other anti-competitive practices, such as tying or bundling, may prove an effective mechanism of retaining dominant market positions even if smaller, AI-focussed firms manage to compete at the foundation level.

How do generative AI models compete?

Developers select generative AI models to provide specific capabilities after balancing various characteristics of the model, including performance and cost. Accordingly, generative AI models compete along multiple different dimensions, which may include any or none of the following:

1. Parameters: Parameters are the internal variables that machine learning models adjust during their training process to improve their ability to make accurate predictions.[10] As the quantity of parameters operating in a model generally equates to the complexity of the tasks the model can complete,[11] parameter growth is often considered the primary method through which to outpace competing models (though an excess of parameters can cause issues too). AI companies have poured resources into upscaling the parameters in their models, as can be observed in the exponential growth in the number of parameters included in OpenAI's Chap GPT model since its launch in June 2018:
    a. GPT-1: 117 million parameters
    b. GPT-2: 1.5 billion parameters
    c. GPT-3: 175 billion parameters
    d. GPT-4: Estimated 170 trillion parameters, based on a backwards extrapolation of the model's performance capabilities.
2. Efficiency: By design, SLMs have comparatively lower parameter counts than their larger counterparts. For example, Microsoft Phi-2 SLM has only 2.7 billion parameters, a quantity of data equivalent to a 5GB file. These models compete on qualitative metrics, seeking to efficiently generate accurate output based on targeted training data (what Microsoft calls "textbook-quality" data).[12] Successful SLMs offer a comparatively resource-light model that can compete with the accuracy of LLMs, particularly when trained and operated on a delimited range of subjects.[13]
3. Agentic architecture: Models' ability to interface with other models is increasingly seen as a key frontier of competitive innovation. These systems are known as agentic systems and may operate within an interconnected web of AI components working in concert or as an independent model capable of reviewing and verifying other models' output. Agentic systems underpin complex automated processes (e.g., self-driving cars) and offer an AI-driven pathway to ensuring output accuracy and combatting the spread of misinformation.
4. Accuracy and reliability: Naturally, a core metric through which an AI model can distinguish itself against competitors is its reliability, with develops seeking to distinguish their models through:
    a. Alignment and model steering through reinforcement-learning, human feedback (RLHF)[14]
    b. Enhancing retrieval-augmented generation (RAG)
    c. New innovations in hallucination detection via Named Entity Recognition (NER)
5. Model relevance: As more and more models appear and the performance gap narrows, we are seeing a change in terms of pure model development as investment shifts into model reasoning, model specificity and model transportability – i.e. the model itself is becoming commoditised and the value is shifting to how and where a model can be used.

Can Australia compete in the GenAI market?

To date, generative AI remains largely synonymous with Silicon Valley. Australia has yet to see a viable foundation model developed domestically and generative AI uptake amongst Australians is reportedly low compared to other nations in

the region.[15] Would-be Australian startups likely face difficulty in obtaining interest from venture capitalists given Australia's underdeveloped AI sector and unfriendly regulatory apparatus.

In particular, any effort to promote the development of generative AI in Australia will need to start with reforms to Australia's data regime. The current absence of a data mining exemption to Australia's copyright laws exposes developers to the risk of copyright infringement when training their models. In its March 2024 response to the Attorney-General's Copyright Enforcement Review Issues Paper, Google flagged that

> "The lack of such copyright flexibilities means that investment in and development of AI and machine-learning technologies is happening and will continue to happen overseas. AI-powered products and services are being created in other countries with more innovation-focused copyright frameworks, such as the US, Singapore, and Japan, and then exported to Australia for use by Australian consumers and businesses. Without these discrete exceptions, Australia risks only ever being an importer of certain kinds of technologies."[16]

This of course raises other issues for which there is little consensus, including about the role of IP law, the legitimacy of a data mining exemption, the value protecting copyright and the moral rights of authors and creators. [17]

Comparing Competition in GenAI technologies and other online markets

In light of the above, there are some notable differences between the competitive characteristics of generative AI technologies and other online technologies that have drawn the attention of competition regulators in recent years. In particular:

1. The integration of a large variety of generative AI models into existing and new applications suggests that its current use is as a general-purpose input or technology that can be used to increase the competitiveness of a product or service in a particular market (see, e.g. the introduction of generative AI chatbots across a range of different online platforms). This is reflected in common pricing and cost structures of large proprietary generative AI models, where enterprise users are charged a fee for access and the model provider incurs a significant cost per query.

2. Unlike multi-sided platforms with strong direct and indirect network effects that occur when the value of a product, service or platform depends on the number of buyers, sellers or users who use it, generative AI technologies do not exhibit strong network effects. That is, a generative AI model does not generate additional value to one user simply because another user is using it, unlike a social network where an increase in users increases the value of the network for all users.

3. Similarly, the same data feedback loop is not present as in advertising-funded online markets such as general search, where newer entrants have struggled to challenge Google's decades-long dominance due to inadequate access to user query data.[18] In contrast, the user prompts and requests made to a generative AI model are generally not used to re-train the model, as initial model training process is not subject to constant improvements on the basis of user queries in the same way as a search algorithm is.

Competition concerns in the Generative AI supply chain

At this stage of its life cycle, generative AI technologies across different markets are also benefiting from a rapid pace of innovation, large scale of investment across the supply chain, and a high number of new entrants. Nevertheless, competitive concerns have emerged regarding aspects of the generative AI supply chain. These include, in the upstream development stages, possible barriers to entry relating to:

1. *Access to Data*: Data is an important input at multiple stages of the generative AI model development process, including during: (i) pre-training, where a dataset is compiled to build the model's knowledge; (ii) fine-tuning, where a foundation model's accuracy is improved through dedicated training such as reinforcement learning with human feedback (RLHF); and (iii) grounding, where a dataset can be used as a reference point when responding to queries in real time.

The data requirements of AI models will vary depending on the size and type of model being trained and the data source being used. Recognising the value of online data, some content-hosting platforms have begun to resist efforts to scrape their data, with news organisations[19] and social media companies such as Reddit[20] and X[21] (formerly Twitter) limiting access to their data. While there are also many publicly available sources of data as well as datasets for training models, this trend may increase barriers to entry for those developers who may require such proprietary data for their model development.

2. *Access to computing power*: Not all AI model developers have the ability, capital, or interest to build their own AI-training data centre infrastructure. This means that many AI start-ups require access to the AI-ready cloud computing infrastructure of other providers, such as Amazon, Microsoft, Google, IBM and Oracle among others.

3. *Access to Talent*: Another barrier to entry may be raised by access to talent: "Developing a generative model requires a significant engineering and research workforce with particular—and relatively rare—skillsets, as well as a deep understanding of machine learning, natural language processing, and computer vision."[22] This competitive resourcing market favours firms able to offer attractive terms (e.g., higher pay and benefits, high equity stake in startup, creative freedom) to attract and retain talent.

Where to from here?

Despite the recent acceleration in generative AI technologies, it remains too early to understand the competitive landscape taking shape and to identify where the market may ultimately be warped. The economies of scale currently operating in the LLM space may prove too onerous to facilitate healthy competition, with only a handful of firms surviving the current boom. Conversely, LLMs may quickly prove antiquated and cumbersome, with subject-matter-specific SLMs that can readily be produced by wide array of firms becoming the dominant AI product. While the technology and related market continue to evolve, careful attention will need to be given to the harms to consumers that arise from the use of generative AI and whether existing consumer and privacy laws are sufficient to combat those harms.

---

[1] We leave for now the the well-reported race to an all-encompassing 'artificial general intelligence': https://www.nytimes.com/2024/06/04/technology/openai-culture-whistleblowers.html, https://www.theverge.com/2024/1/18/24042354/mark-zuckerberg-meta-agi-reorg-interview, https://www.truthdig.com/articles/the-madness-of-the-race-to-build-artificial-general-intelligence/

[2] https://www.ftc.gov/legal-library/browse/joint-statement-competition-generative-ai-foundation-models-ai-products.

[3] https://www.gov.uk/cma-cases/ai-foundation-models-initial-review.

[4] https://www.accc.gov.au/inquiries-and-consultations/digital-platform-services-inquiry-2020-25/march-2025-final-report.

[5] See CMA Technical Update Paper, para 24.

[6] https://www.nasdaq.com/articles/nvidia-dominating-artificial-intelligence-chip-market-apple-has-been-securing-supply

[7] See HuggingFace Models Library (link available here). Further details of the number of open LLMs available can be found at HuggingFace Open LLM Leaderboard (link available here).

[8] As of December 2023, a total of approximately 190,000 transformer-based models were listed on Hugging Face.

[9] https://news.microsoft.com/de-ch/2024/02/27/microsofts-ai-access-principles-our-commitments-to-promote-innovation-and-competition-in-the-new-ai-economy/

[10] https://ourworldindata.org/grapher/artificial-intelligence-parameter-count

[11] See https://medium.com/@sanjeeva.bora/understanding-parameters-in-genai-models-design-use-cases-and-significance-9a0c03902dff.

[12] https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

[13] See, eg Microsoft's Orca-Math model, a specialised SLM designed to solve maths problems: https://www.microsoft.com/en-us/research/blog/orca-math-demonstrating-the-potential-of-slms-with-model-specialization/

[14] See further: https://www.unimelb.edu.au/caide/caide-law-2024/caide-law-past-events/demystifying-generative-ai

[15] https://www.deloitte.com/au/en/services/economics/blogs/generative-ai-she-be-right-approach-australia.html

[16] https://consultations.ag.gov.au/rights-and-protections/copyright-enforcement-review/consultation/view_respondent?sort=excerpt&order=ascending&uuId=665660477

[17] See also Generative AI & IP Issues Paper (2024).

[18] See the explanation of this advantage in the recent antitrust decision United States vs Google (2024): "Because many users simply stick to searching with the default, Google receives billions of queries every day through those access points. Google derives extraordinary volumes of

user data from such searches. It then uses that information to improve search quality. Google so values such data that, absent a user-initiated change, it stores 18 months-worth of a user's search history and activity."

[19] https://www.theguardian.com/technology/2023/aug/25/new-york-times-cnn-and-abc-block-openais-gptbot-web-crawler-from-scraping-content

[20] https://www.platformer.news/reddit-goes-dark/

[21] https://www.businessinsider.com/elon-musk-openai-twitter-data-pay-dispute-2023-4

[22] https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns

THE UNIVERSITY OF
MELBOURNE

Centre for AI and Digital Ethics

Level 8
Melbourne Connect
700 Swanston Street
Carlton 3053

unimelb.edu.au/caide