"Generative AI is the most important technology of the 21st century. It has the potential to solve some of the world's biggest problems, such as climate change and disease."
Yann LeCun, Founding Director of the NYU Centre for Data Science

Before we delve into Generative AI (GenAI), let's consider the challenges faced by the poor Optus salesperson after its second 'event' in a year.

What do you tell customers?

Almost nothing would suffice. Not due to the catastrophic network failure, but because of Optus's response, or rather its silence! Even the Minister, jumped on radio (note) as soon as she could.

Unknowingly, Optus is providing invaluable but painful lessons on crisis management, albeit potentially expensive ones if you are a customer. And this one wasn't about cybersecurity.

This month we discuss Generative AI (GenAI) and and what it means to business. Two takeaways:

- GenAI is the most important transformational technology since the internet
- The rate of change occurring in GenAI is faster than any transformational technology in history

Previous Newsletters, including this one, are available on our site in pdf HERE

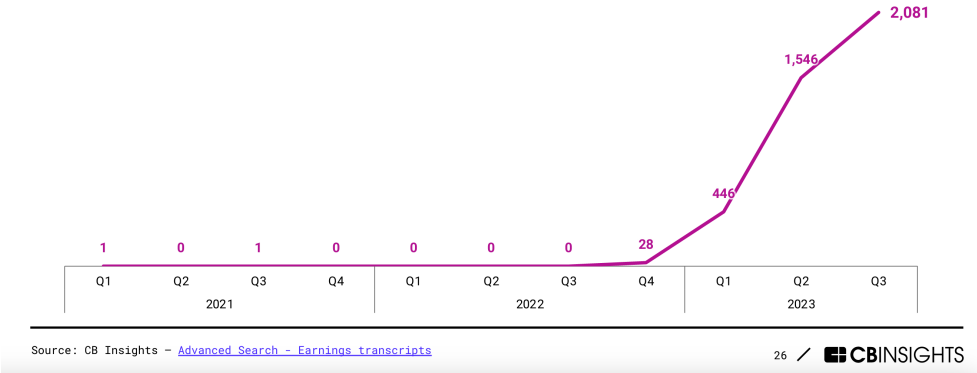
Enterprises interest in GenAI surges

Sixty-one percent of respondents believe they have a maximum of one year to implement an AI strategy before their organisation begins to incur significant negative business impact.

Cisco research highlighting seismic gap in companies' preparedness for AI, November 2023 LINK

While you can undoubtedly find out more about GenAI for yourself using GenAI, the following provides an overview of the key aspects of the technology and market dynamics.

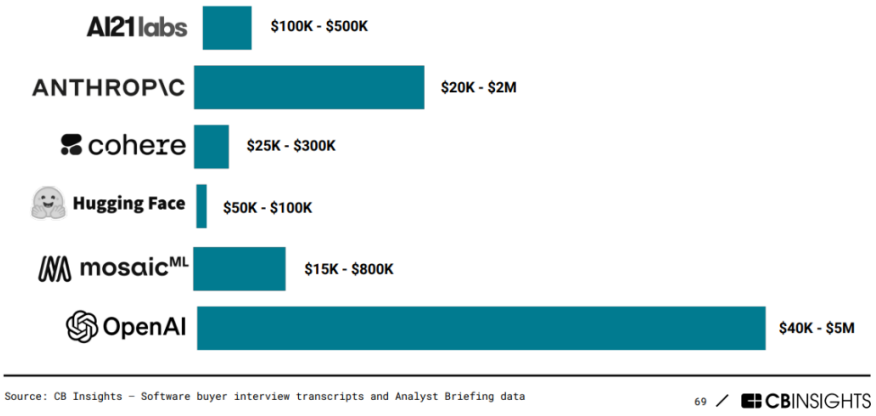
INTEREST: The astonishing capability of GPT 3.5, followed by its global availability via ChatGPT in November 2022, was like the lightening before the thunder that reverberated globally. The graph below shows how business interest in GenAI has skyrocketed in just a few months.



Almost overnight, executive interest in GenAI skyrockets and companies feel pressured to react.
Earnings call mentions of "Generative AI" (as of 30th Sept 2023)

URGENCY: Sixty-one percent of respondents to recent Cisco research believe they have a maximum of one year to implement an AI strategy before their organization begins to incur significant negative business impact LINK.

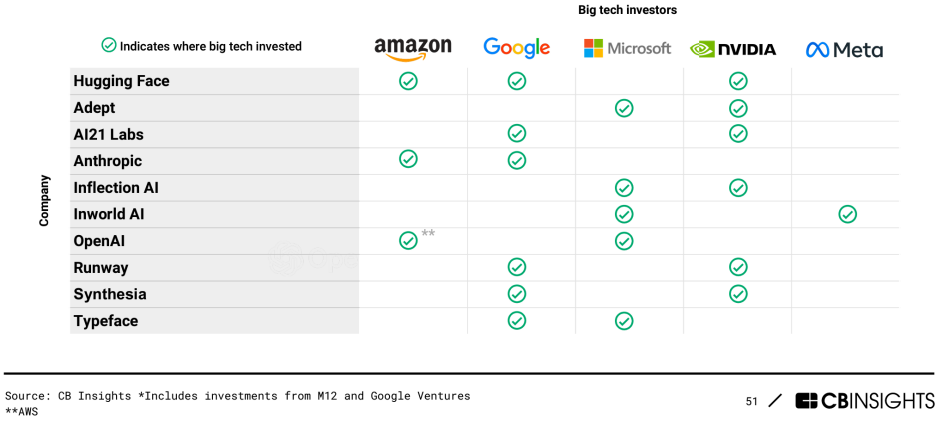
SPENDING: Enterprises are spending big with large language model (LLM) vendors to deploy AI models.



Enterprises are spending millions with LLM developers (Annual spend ranges displayed)

LLM VENDORS: While OpenAI has a clear lead, vendors are competing on multiple fronts to become the go-to developer. Out of the sixteen new AI unicorns in 2023 so far, eleven are GenAI companies. Most of the venture funding is currently being invested into AI infrastructure, critical for the ever-increasing scale of the models. This also explains the next evolution of cloud (sometimes referred to as "Cloud 2.0") and Nvidia's market cap passing US$1 Trillion as it has extended such a wide lead on AI infrastructure.

As the chart below illustrates, GenAI is a new battleground for big tech, with overlapping alliances and commitments into the billions.

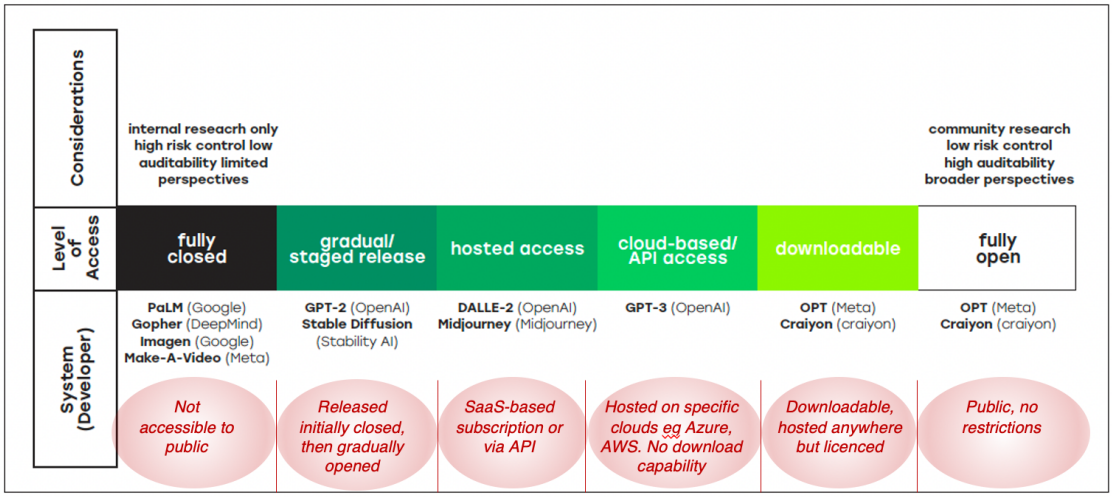


GenAI companies with two or more big tech investors (as of 30th Sept 2023)

CONCERNS: While businesses recognise the potential benefits, given GenAI's relative immaturity, there are justifiable concerns. These include whether they have sufficient internal expertise, selecting the right path and options, availability of relevant data, potential exposure of internal data, regulation, understanding the risks and many more.

OPEN-SOURCE VS PROPRIETARY MODELS: There are already a range of LLM options ranging from closed to open source. Open-source models can be downloaded and used by anyone who has the knowledge to host them on their own hardware. As featured in last month's newsletter, Hugging Face, founded in 2016, hosts hundreds of thousands of AI models. It serves as a GitHub-like hub for AI code repositories, models, datasets, and web apps. LINK.

There is a spectrum of LLMs (shown below) with varying degrees of access ranging from closed models like PaLM (Google), to open models, like OPT (Meta), with no restrictions.



As systems become more open, they are more auditable, but more difficult to control for risks.
Source: Adapted from Tesserent "AI in the Enterprise", November 2023

EVALUATION CRITERIA: Key evaluation criteria for buyers include:

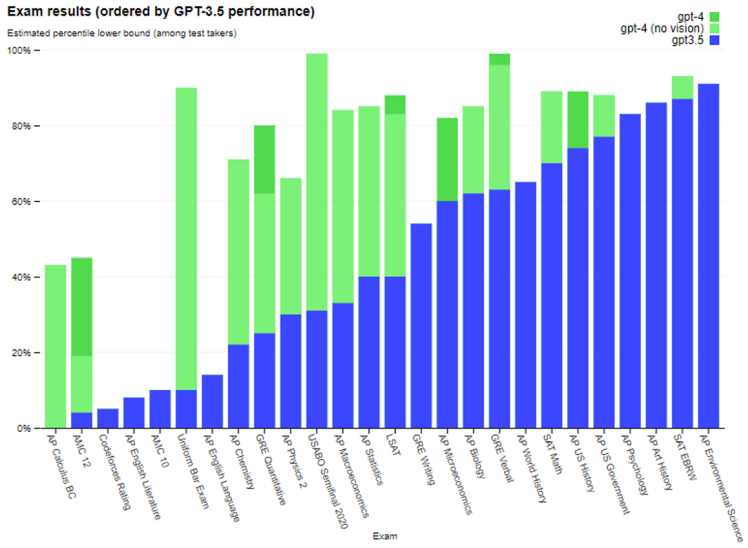
- Safety & compliance
- Accuracy & quality
- Customisation
- Pricing & deployment
- Token limits
- Privacy and data handling

But other factors also play into the buying decisions for foundation models.

PACE: The rate of change is breathtaking - never before has the industry witnessed such a rapid rate of change. A client recently produced an excellent paper on LLM evaluation. Within a week it was dated. The authors are leaders within their business so one wonders what this means for everyone else trying to keep up. What does it mean for your business?

"These models will keep getting better, so the field is fast changing"
Alphabet CEO Sundar Pichai Q1'23 earnings call

CAPABILITY: The number of parameters feeding the LLMs has risen from 175 billion (GPT-3.5) to over 100 trillion (GPT-4.0) within the space of a year. The results are astounding. As shown below, GPT-4.0 performance is outperforming humans in tests from bar exams to environmental sciences.



GPT-4 becomes OpenAI's most powerful model yet, crushing human exams

Don't assume that the larger the model the better. In still largely exploratory work, Microsoft researchers showed that when small language models (SLMs) are trained with very specialised and curated datasets, they can rival models which are 50x larger. They also find that these models' neurons are more interpretable. LINK

CODE GENERATION: Generative AI has made significant progress in code development. The leader in terms of coding abilities is unsurprisingly GPT-4, with Code Interpreter or now, Advance Data Analysis, leaving users in awe. Open alternatives like WizardLM's WizardCoder-34B and Unnatural CodeLLaMa hold up with ChatGPT in coding benchmarks, but their performance in production is still TBD.

APPLICATIONS/VERTICAL AI: While the initial focus in GenAI is on capabilities, platforms and tooling, interest is expected to quickly shift to its application. There will be several 'horizontal' use cases across multiple sectors, however use of GenAI for vertical-specific applications holds much promise. Sectors like FSI, Healthcare and Retail are progressively exploring ways to drive growth, improve customer experience and reduce cost and risk. The advent of specialised "Copilots" for doctors, financial advisors, pharmacists, architects, lawyers and many more roles, is imminent.

Google's Med-PaLM was the first model to exceed a "passing" score on the US Medical Licencing Examination (USMLE). A year later, in a ranking study on 1,066 consumer medical questions, Med-PaLM 2 answers were preferred over physician answers by a panel of physicians across eight of nine axes in their evaluation framework.

The Future

The future is already here - just look at OpenAI's first developers' conference on 6th November. It provides a front row seat for what to expect in 2024:

- ChatGPT user milestone: There are over 100 million weekly active users and 92% of the Fortune 500 companies using its product. More than 2 million developers are building solutions through its API.
- GPT-4 Turbo: an improved version of the popular GPT-4 model that is more powerful, lower cost (~50%) and has a knowledge cut-off of April 2023, compared to GPT-4's September 2021 cutoff.
- Build your own GPTs: Users will be able to build these GPTs ('agents' or 'bots') just through prompts without needing to know any coding. Enterprise customers can make internal-only GPTs built on top of the company's knowledge bases like emails to import external information.
- New GPT store for user-created AI bots: Users can publish these GPTs to a store. It will initially have creations from "verified builders". Developers will get paid for published GPTs.
- New API to let devs build assistants: New Assistant APIs let developers build their own "agent-like experiences". Developers can make agents that retrieve outside knowledge or call programming functions for a specific action. Use cases range include coding assistants and AI-powered vacation planners.
- DALL-E 3 API: OpenAI's text-to-image model DALL-E 3 is now available through API with in-built moderation tools.
- New text-to-speech APIs: Called Audio API comes with six preset voices.
- A promise to defend businesses from copyright claims: Copyright Shield promises to protect businesses using OpenAI's products from copyright claims. The company will pay legal fees if customers using "generally available" OpenAI's developer platform and ChatGPT Enterprise face IP lawsuits against content created by OpenAI's tools.

In summary, several important new products, simplifying and incentivising third-party application ("agent") development (on GPT) and significantly lower prices.

Generative AI is profoundly important, it is happening fast, it is now. If you would like a quick assessment of where your business stands in terms of AI maturity, you may want to try "Cisco's AI Readiness Assessment" LINK.

Stay connected.

Kevin

PS. You may want to try my GenAI bot HERE. It took less than 15 minutes to set up,