Generative AI

solutions for innovation

February 2024

table of contents

acknowledgements

Author

Ying Shaowei
Chief Scientist, NCS

As the head of its Technology Centre of Excellence, Shaowei drives NCS' innovation strategy in the intelligence-led era. He actively contributes to ongoing NCS efforts to develop AI-powered industry business solutions and accelerators, providing executive sponsorship and directing investment in high-impact programmes.

Shaowei also shapes the innovation agenda with NCS' strategic accounts and key technology partners, including local universities, national research institutes, and corporate R&D labs.

Additionally, he plays a crucial role in developing NCS' voice on emerging technologies, articulating long-term perspectives on critical technology trends that influence businesses.

Contributors

NEXT

Wynthia Goh
Senior Partner, NEXT Centre Management

Dr. Sunil Sivadas
Practice Lead, NEXT Gen Tech

Ng Chong Yang
Lead Product Innovator, NEXT Open Innovation

Royston Bok
Senior Director, NEXT Growth

Juan Miguel Jimeno
Senior Research Scientist, NEXT Gen Tech

Josey Mathew
Senior Research Scientist, NEXT Gen Tech

Cyber

Lim Hoon Wei
Principal Director, Cyber Special Ops and R&D

Daniel Ong
Senior Consultant, Cyber Management

Tech CoE

Vijay Nayudu
AI Business Lead, Tech CoE

Nikita Atkins
AI Business Lead, Tech CoE

Samin Batra
Research Scientist, Tech CoE

Temo Anda
Senior Data Scientist, Data Spark

Editorial support

Georgia Swanson
Director of Strategy & Growth, NCS NEXT Global Innovation

Mark Addy
Insights Lead, NCS NEXT, Global Innovation

Deirdre Chong
Visual Designer, NEXT Open Innovation

foreword

As NCS Chief Scientist, I have the privilege of welcoming and hosting many distinguished visitors (senior business executives and government officials) to NCS Tesseract, our innovation centre that showcases emerging technologies. Since the opening of the Tesseract in May 2023, the topic of artificial intelligence (AI) has dominated conversations at the Tesseract.

Our visitors spend a disproportionate amount of time during their visits viewing our AI showcases, sharing their various business pain points and needs, and discussing the realm of possibilities for AI. AI in general, and Generative AI (GenAI) in particular, has been developing at such an astonishing pace that it has become difficult for managers to determine how best to evaluate and implement the technology, and how to keep current on the evolving changes to AI.

These discussions have inspired us to write this collection of articles.

Shaowei Ying,
Chief Scientist, NCS

NCS has been investing extensively in the human and technological capabilities required to both capitalise on the benefits and minimise the risks of AI integration into our own solutions. Through our shared experience conducting research, developing and testing models, collaborating with industry peers and government entities, and delivering GenAI solutions, we've developed a level of expertise and understanding of GenAI that we share with you in this report.

Whether you are just starting to understand and evaluate how GenAI might benefit your organisation, or you are well on your way in implementing and using this revolutionary technology, we hope that you find the information contained in this report useful.

NCS would welcome the opportunity to further explain how we are using GenAI and to collaborate with you on your GenAI journey.

introduction

Generative artificial intelligence (GenAI) is a rapidly advancing field within artificial intelligence that leverages sophisticated machine learning techniques to produce and generate diverse forms of content. This can encompass a broad spectrum of outputs, from the generation of coherent and contextually relevant text such as articles, creative writing, and dialogues, to creating visually striking and unique imagery, design layouts, and videos, to the creation of audio content.

However, the excitement surrounding GenAI is not, and should not be, only about the current state of the technology, but also its future potential. Next-generation innovations in this field are set to deepen the integration of GenAI into everyday applications and business processes, making them more personalised, context-aware, and interactive. Upcoming innovations may allow for more advanced content adaptation, real-time generative capabilities, and improved creative collaboration between AI tools and humans. GenAI is transforming how we approach problem-solving, creativity, and data analysis. It's not just about automating tasks, bu about augmenting our capabilities, enhancing creativity, generating novel insights, and enabling new forms of communication and interaction.

NCS is exploring emerging GenAI capabilities to continuously improve the creativity, relevance, and impact of all NCS solutions. Our AI initiatives currently focus on enhancing co-creativity (where the AI works in tandem with human users to generate content), improving the contextual understanding of AI tools in order to generate more relevant output, and expanding accessibility and usability across a wider range of applications. By staying at the forefront of AI innovation, we aim to unlock unprecedented opportunities for creativity, efficiency, and value creation.

get involved

Given the rapid pace of GenAI development, continuous learning, exploration, and adaptation will be key to the successful utilisation and adoption of this new technology.

As we journey into this exciting future, we invite you to stay engaged with NCS. Follow our latest work in Generative AI, learn from NCS thought leaders, contribute your unique insights, and be a part of the future we're building together.

AI everywhere, skin-deep

Ying Shaowei

AI everywhere, skin-deep

Ying Shaowei

Artificial intelligence (AI) has rapidly evolved from a theoretical concept into a ubiquitous technology that touches nearly every industry and aspect of our life. However, despite its pervasiveness in conversation and its potential to revolutionise businesses, AI's integration into operations is superficial. The reality is that while AI appears to be everywhere, its impact, depth of integration, and economic benefits are still emerging and unevenly distributed.

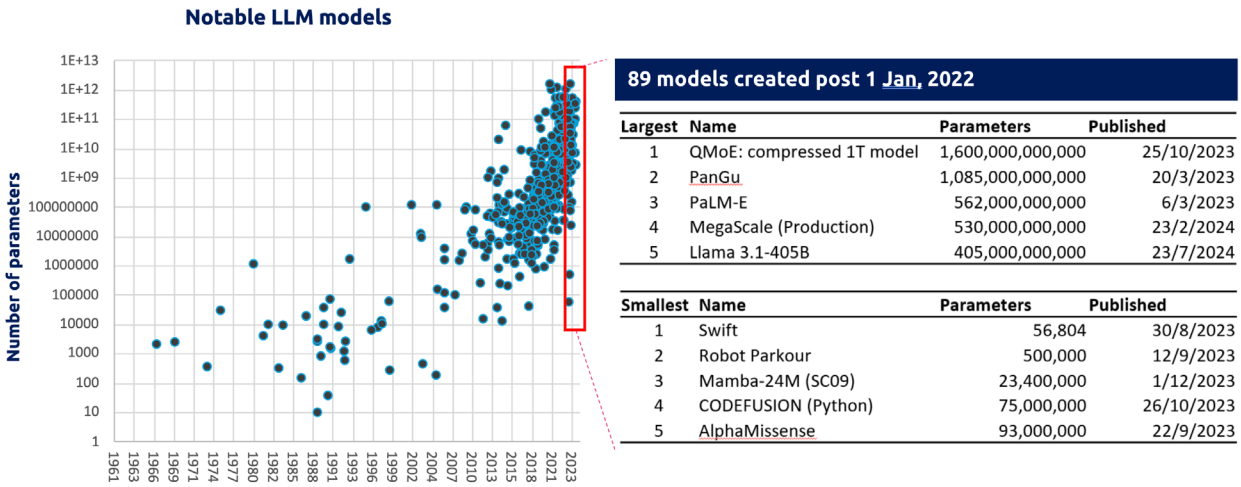
This article provides an updated perspective on AI technology trends and state of AI adoption.

Technology trends: the current landscape

1) Size is not everything

The early wave of AI advancements focused on developing large, general-purpose models, such as the initial large language models (LLMs). These frontier models are becoming increasingly powerful, now incorporating multimodal capabilities that allow them to process and generate both text and images. However, as these models grow in size and complexity, they also bring significant sustainability challenges. The vast amounts of energy required to train and deploy these massive models result in substantial carbon footprints, raising concerns about their long-term environmental impact. GPT-3 which has 175 billion parameters is estimated to have produced ~550 metric tons of CO_2 equivalent emissions (Stanford Institute for Human-Centered Artificial Intelligence (HAI), 2023). To put it in perspective, this amount of CO_2 is roughly equivalent to flying from Singapore to New York USA and back 140 times.

In response to these challenges, a parallel movement is gaining traction, emphasising the development of smaller, task-specific models (Figure 1). These models are designed to perform well on specific tasks with greater efficiency, consuming far less energy and reducing environmental impact. Despite their smaller size, these models often rival larger ones in their respective domains, demonstrating that in AI, bigger isn't always better. The shift towards more sustainable, use-case-driven solutions highlights a growing recognition of the need for AI to be both powerful and environmentally responsible.



89 models created post 1 Jan, 2022

Largest	Name	Parameters	Published
1	QMoE: compressed 1T model	1,600,000,000,000	25/10/2023
2	PanGu	1,085,000,000,000	20/3/2023
3	PaLM-E	562,000,000,000	6/3/2023
4	MegaScale (Production)	530,000,000,000	23/2/2024
5	Llama 3.1-405B	405,000,000,000	23/7/2024

Smallest	Name	Parameters	Published
1	Swift	56,804	30/8/2023
2	Robot Parkour	500,000	12/9/2023
3	Mamba-24M (SC09)	23,400,000	1/12/2023
4	CODEFUSION (Python)	75,000,000	26/10/2023
5	AlphaMissense	93,000,000	22/9/2023

Figure 1: Notable AI models by timeline and the number of weights or trainable parameters. (Epoch AI, n.d.)

2) The AI fork: open vs. closed

A significant shift in AI technology is the rise of open models, which are publicly accessible through open-source, open-weight parameters or both, enabling broader use and further development by community. These open models are increasingly rivalling proprietary, closed models.

Historically, major tech companies developed and dominated the AI landscape with their proprietary models, limiting access to the most advanced capabilities. However, the emergence of powerful open-source models is democratising access to cutting-edge AI, enabling a broader range of organisations to leverage AI technologies. Figure 2 shows how open-weight models, particularly with the release of Llama3.1 in July 2024, are closing the gap with closed-source ones.

These open models offer cost advantages and foster innovation through community-driven development. As more businesses and researchers contribute to these models, they rapidly improve in performance and usability. The growing adoption of open-source models represents a significant shift in how AI technologies are developed and deployed. We can expect more open models challenging the dominance of closed, proprietary systems and offering new opportunities for businesses to innovate.

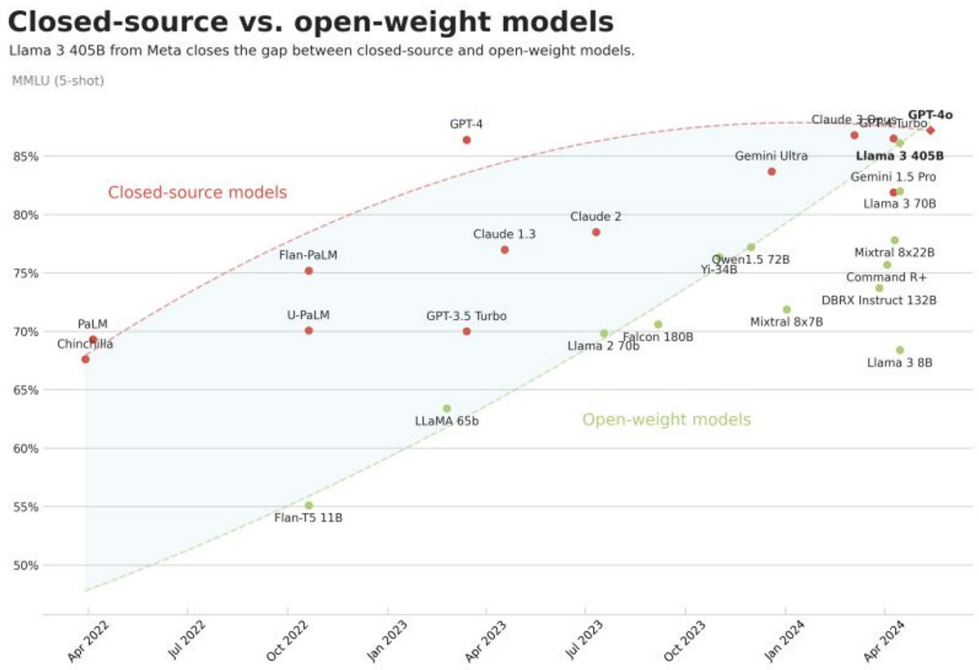


Figure 2: Open-weight models closing the gap with closed models (Labonne, 2024)

3) From RAG to riches

Retrieval augmented generation (RAG) has emerged as a cornerstone of enterprise use of Generative AI. As shown in Figure 3, RAG combines retrieval-based systems where information is pulled from the organisation's database of existing knowledge, with Generative AI models that create new content based on that information. This hybrid approach ensures that the risks of AI model "hallucination" are reduced and that the AI outputs are not only accurate but also contextually relevant, providing tailored insights for business decisions. As businesses increasingly rely on AI, RAG enhances the quality and reliability of AI applications, underscoring the need for sophisticated data management and well-structured strategies to unlock AI's full potential.

In the near future, we can anticipate a significant increase in tools and services leveraging RAG to enable enterprises to effectively mine their proprietary and third-party data using Generative AI. RAG is likely to be viewed as a more accessible and cost-effective alternative to fine-tuning LLMs.

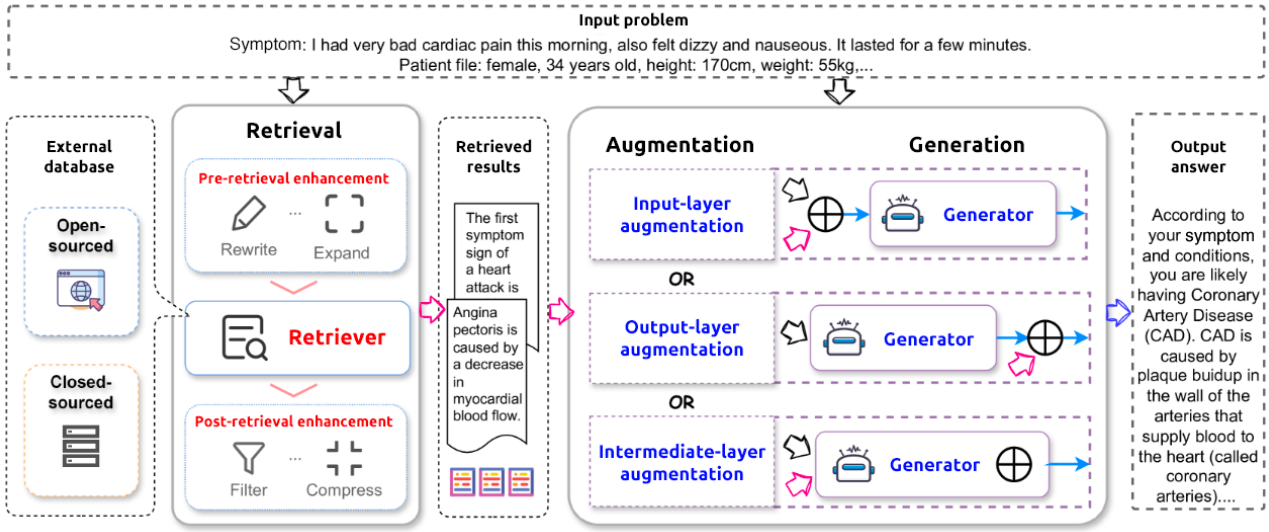


Figure 3. The basic retrieval augmented large language models (RA-LLMs) framework for a specific QA task (Ding, et al., 2024). RAG is continuing to evolve rapidly around 4 areas, namely (1) the RAG Framework/Pipeline, which integrates retrieval mechanisms with generative models to boost the accuracy and relevance of AI-generated outputs; (2) RAG Learning, focused on optimising the interaction between retrieval and generative components through strategic training; (3) Retriever Learning, which involves fine-tuning algorithms to accurately identify and retrieve relevant information from databases; and (4) Pre-/Post-Retrieval Techniques, employed to refine the quality of retrieved data and enhance its effectiveness in generating precise and contextually appropriate AI outputs.

Ecosystem shifts: the influence of giants and the local context

4) Bottleneck in GPUs: giants lead the race

With a market share north of 85%, NVIDIA is the most dominant AI hardware vendor, providing the advanced computing chips or GPUs crucial for training and deploying the LLMs and other foundation models. In 2023, Microsoft and Meta emerged as significant buyers of NVIDIA's H100 GPUs, spending $9 billion combined and acquiring around 150,000 chips each. Additionally, Google and Amazon purchased approximately 50,000 chips each. Tesla, another large customer, bought 15,000 chips, with plans to increase this number significantly. These figures highlight the substantial investments by these tech giants in AI infrastructure, contributing nearly 40% of NVIDIA's revenue (Tremayne-Pengelly, 2024).

The rapid adoption of AI by the tech giants has created a bottleneck in the supply of GPUs. These giants have the resources to secure these components, leaving smaller players struggling to keep up. Analysts are predicting the GPU market will grow at a CAGR of 34% over the next decade (Figure 4). With the current industry dynamic, it will not be surprising that this growth be accompanied by an increasingly uneven playing field where only the most resourceful companies can thrive.

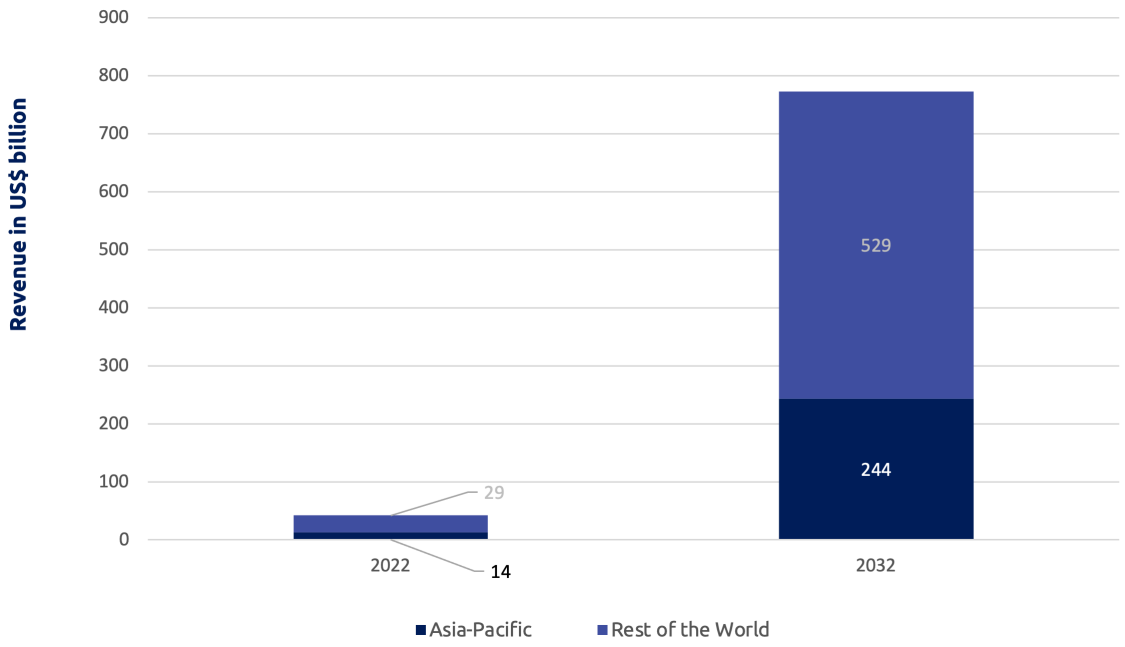
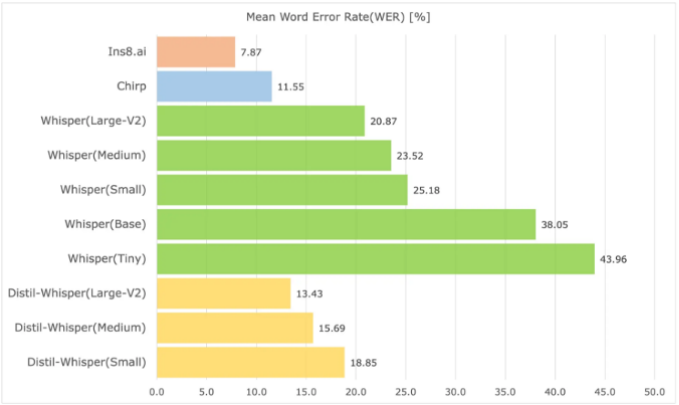


Figure 4. GPU market size in 2022-2032 (Precedence Research, 2023)

5) Localisation: tailoring AI to local markets

Localisation is becoming increasingly important in AI deployment. To be truly effective, AI systems must consider the specific needs and conditions of local markets. This has led to the rise of new AI players who offer industry-focused solutions tailored to local contexts, providing businesses with more choices in navigating AI adoption.

Figure 5 shows the ongoing work that is done by NCS to build a conversational AI system that could understand the Singaporean-accented speech with its use of borrowed words from other languages and many references to local places and events. It is an example of how localisation can improve the performance of AI systems.



Organisation	Model	Parameters(M)
NCS Pte Ltd	Ins8.ai	120
Google	Chirp	2,000
OpenAI	Large-V2	1,550
OpenAI	Medium	769
OpenAI	Small	244
OpenAI	Base	74
OpenAI	Tiny	39
HuggingFace	Distil-Large-V2	756
HuggingFace	Distil-Medium	394
HuggingFace	Distil-Small	166

Figure 5. Relative performance of AI systems for speech-to-text (STT) transcription of Singlish. Most globally trending STT engines (such as those from OpenAI and Google) are not optimised for understanding languages and their dialects spoken by small populations and communities. AI models trained with local datasets can perform better with fewer parameters and smaller computational footprint, as this example of Ins8.ai by NCS demonstrates.

6) Global AI landscape: not all that glitters is Western

While the US is the clear leader in the source of top AI models, China dominates AI patents. In 2022, China's patenting activities significantly outstripped the US, accounting for 61% of global AI patent origins compared to US' 21% (Stanford Institute for Human-Centered Artificial Intelligence (HAI), 2023). Notably, in Hugging Face's second leaderboard that ranks open LLMs against more challenging variety of tasks, Alibaba's 千问 (Qianwen) models featured prominently among the top-ranked open models on Hugging Face leaderboard.

The notion of AI leadership being confined to the West is misplaced. China's rapid development and official approval of adoption of LLMs, urging both citizens and enterprises to embrace AI on a large scale, highlights the prospect that China will continue to be a hotbed for many AI innovations to come and can provide businesses and governments with meaningful alternative AI solutions.

SN	Model	Description					Performance
		Country of origin	Type	Architecture	#Params (B)	Submission date	MMLU-PRO
1	Qwen/Qwen2-72B	China	pretrained	Qwen2ForCausalLM	72	26/6/24	52.56
2	cognitivecomputations/dolphin-2.9.2-qwen2-72b	China	chat models (RLHF, DPO, IFT, ...)	Qwen2ForCausalLM	72	27/6/24	49.52
3	Qwen/Qwen2-72B-Instruct	China	chat models (RLHF, DPO, IFT, ...)	Qwen2ForCausalLM	72	26/6/24	48.92
4	Qwen/Qwen1.5-110B	China	pretrained	Qwen2ForCausalLM	111	13/6/24	48.45
5	CausalLM/34b-beta	USA	fine-tuned on domain-specific datasets	LlamaForCausalLM	34	26/6/24	48.06
6	meta-llama/Meta-Llama-3.1-70B-Instruct	USA	chat models (RLHF, DPO, IFT, ...)	InternLM2ForCausalLM	70	15/8/24	47.88
7	meta-llama/Meta-Llama-3-70B-Instruct	USA	chat models (RLHF, DPO, IFT, ...)	LlamaForCausalLM	70	12/6/24	46.74
8	abacusai/Smaug-Qwen2-72B-Instruct	China	chat models (RLHF, DPO, IFT, ...)	Qwen2ForCausalLM	72	29/7/24	46.56
9	Qwen/Qwen1.5-110B-Chat	China	chat models (RLHF, DPO, IFT, ...)	Qwen2ForCausalLM	111	12/6/24	42.5
10	abacusai/Smaug-Llama-3-70B-Instruct-32K	USA	chat models (RLHF, DPO, IFT, ...)	LlamaForCausalLM	70	6/8/24	41.83
11	01-ai/Yi-1.5-34B-32K	USA	pretrained	LlamaForCausalLM	34	12/6/24	41.21
12	meta-llama/Meta-Llama-3-70B	USA	pretrained	LlamaForCausalLM	70	12/6/24	41.21
13	microsoft/Phi-3-medium-4k-instruct	USA	chat models (RLHF, DPO, IFT, ...)	Phi3ForCausalLM	13	12/6/24	40.84
14	01-ai/Yi-1.5-34B	USA	pretrained	LlamaForCausalLM	34	12/6/24	40.73
15	meta-llama/Meta-Llama-3.1-70B	USA	pretrained	InternLM2ForCausalLM	70	23/7/24	40.6
16	cognitivecomputations/dolphin-2.9.3-Yi-1.5-34B-32k	USA	fine-tuned on domain-specific datasets	LlamaForCausalLM	34	27/7/24	40.34
17	abacusai/Smaug-72B-v0.1	USA	fine-tuned on domain-specific datasets	LlamaForCausalLM	72	12/6/24	40.26
18	HuggingFaceH4/zephyr-orpo-141b-A35b-v0.1	France	chat models (RLHF, DPO, IFT, ...)	MixtralForCausalLM	140	12/6/24	39.85
19	Qwen/Qwen2-57B-A14B-Instruct	China	chat models (RLHF, DPO, IFT, ...)	Qwen2MoeForCausalLM	57	14/8/24	39.73
20	cognitivecomputations/dolphin-2.9.2-Phi-3-Medium	France	chat models (RLHF, DPO, IFT, ...)	MistralForCausalLM	-1	5/8/24	39.5

Figure 6. Hugging Face leaderboard of top-ranked open LLMs. The ranking is by the Massive Multitask Language Understanding Pro (MMLU-PRO) benchmark, which is used for evaluating the performance of LLMs on a broader and more challenging set of tasks compared to the original MMLU, designed to test the latest models' capabilities in handling complex, real-world problems. Alibaba's 千问 models (abbreviated as Qwen) is highlighted in green. (https://open-llm-leaderboard-blog.static.hf.space/dist/index.html)

Superficial integration: the reality of AI adoption

While the technology trends in AI are impressive, a closer examination reveals that many businesses are still engaging with AI at a superficial level. Despite the hype and the fear of missing out (FOMO), AI adoption often lacks depth, driven more by external pressures than by a deep, strategic understanding of how AI can transform operations. According to the 2024 Gartner Hype Cycle Report, "Generative AI is sliding into the trough of disillusionment" (Mearian, 2024).

7) FOMO is catching on

A recent survey by IDC reveals that in the Asia-Pacific region, only 4% of respondents have yet to introduce AI into their operations (Figure 7). This widespread interest in AI is not surprising, given the impressive technological advancements and the continuous hype from technology vendors. Additionally, national initiatives like Singapore's National AI Strategy 2.0 are fuelling this momentum, encouraging businesses to adopt AI to enhance competitiveness across industries.

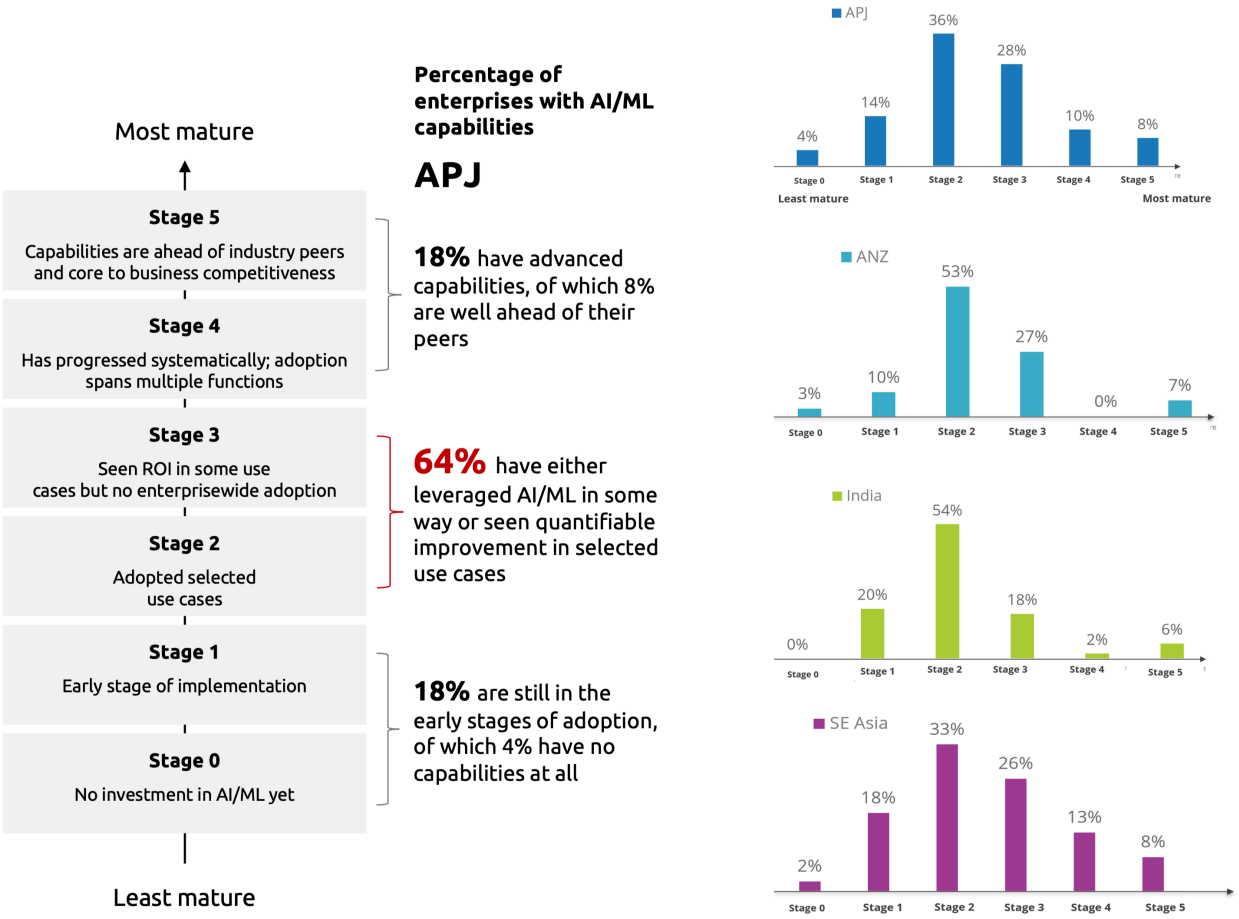


Figure 7. AI maturity in APJ organisations. IDC survey in 2023 shows that the majority of surveyed organisations (82%) are still in early stages of harnessing AI, and have not achieved scaled impact yet. Only a small percentage (4%) have not done anything and are yet to respond to recent advances in AI technology.

8) Reality check

However, as businesses wade into the AI era, they are starting to encounter challenges that temper their enthusiasm. Usability, safety, and enterprise readiness have surfaced as significant concerns. Many AI systems, while powerful, are not yet fully equipped to handle the complexities of real-world applications. This has made businesses more cautious about which use cases to pursue and how deeply to integrate AI into their operations.

Moreover, the lack of skilled personnel to manage and interpret AI systems often leads to suboptimal outcomes. Without the necessary expertise, businesses struggle to fully harness AI's capabilities, resulting in implementations that fall short of expectations. Their use of AI tends to be experimental in nature, or in localised production systems for specific narrow parts of the business.

This state of "AI dust" can be overcome. To fully harness the potential of AI, companies must not only have a well-thought AI adoption strategy that factors in talent and ethics, they must also build the corresponding digital resilience. In other words, they will need a holistic approach of getting IT systems and digital platforms (which host the AI systems) to be ready for the AI era. This holistic approach can be visualised in an AI+DR Impact Matrix (Figure 8), which provides a practical framework to assess an organisation's readiness in AI and digital resilience (DR).

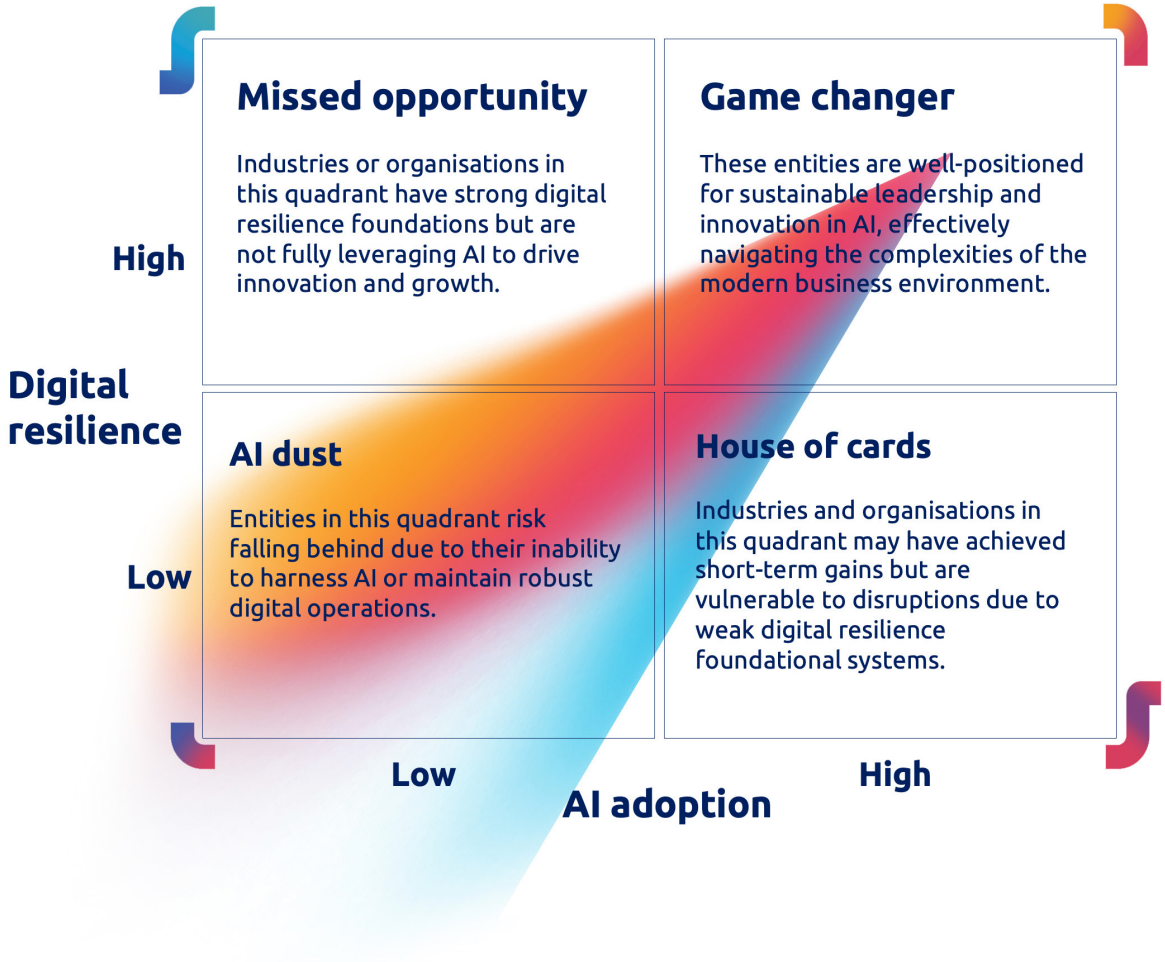


Figure 8. AI+DR matrix. Organisations can use the AI+DR Impact Matrix to engage in strategy discussions, identifying gaps, and prioritising areas for investment not just in AI but also in DR. The matrix serves as a strategic reminder to ensure that as AI adoption increases, corresponding investments in digital resilience are made to maintain stability and security. (NCS, 2024).

9) Regulation rising

As AI technologies become more pervasive, so do concerns around their misuse and potential for abuse. Geopolitical tensions and ethical considerations are driving the development of regulatory measures aimed at controlling the deployment and impact of AI. These regulations, while necessary, add another layer of complexity for businesses trying to navigate the AI landscape. Companies, particularly multinationals, must now consider not only the technical and operational aspects of AI adoption but also the legal and ethical frameworks that govern its use in the markets they operate in.

Not all the AI regulations enacted by countries are restrictive in nature, as shown in Figure 9. In Singapore, the Government has elected to adopt a "twin engine approach to AI regulation" (Yeong, 2024). Rather than implementing broad, sweeping regulations at this time, they have decided to pursue a more balanced approach that equally prioritises the adoption of AI technology and the safeguarding of consumer interests.

Regardless, the rise of AI regulation reflects the growing recognition that AI's influence extends far beyond business operations, impacting society at large. As governments and regulatory bodies introduce new rules and guidelines, businesses will need to adapt quickly to ensure compliance, which could further slow the pace of deep, meaningful AI adoption.

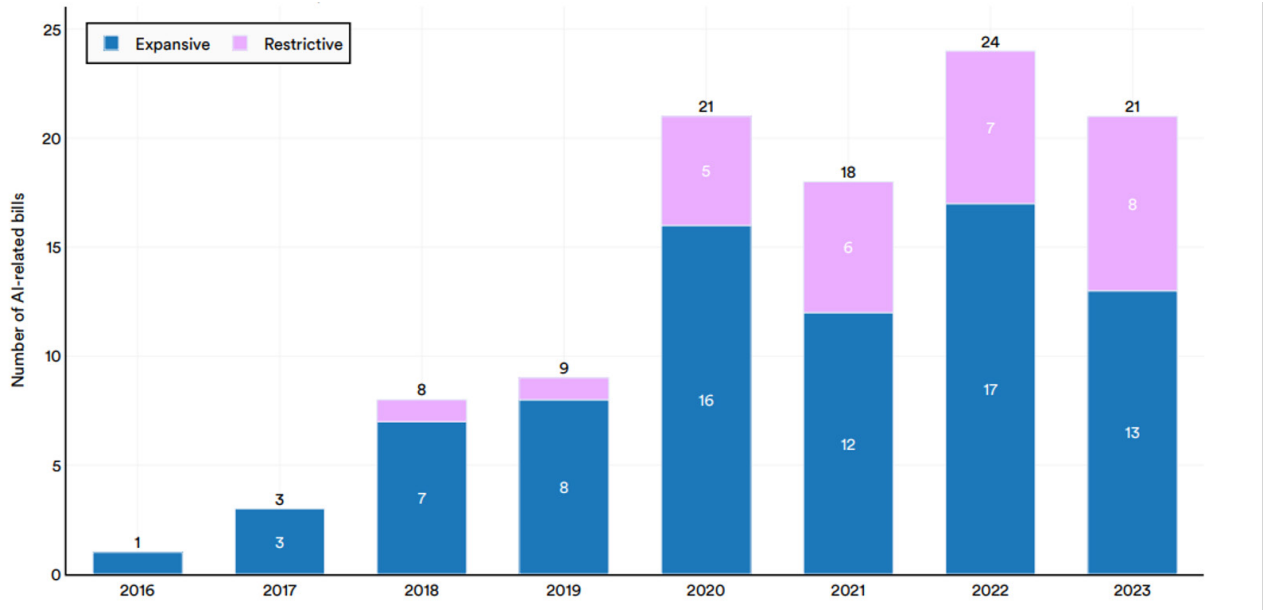


Figure 9. The number of AI-related bills passed into law has seen a dramatic increase since 2019 in the countries studied by Stanford Institute for Human-Centered Artificial Intelligence (HAI) (2023), namely Andorra, Austria, Argentina, Belgium, France, Hungary, Italy, Kazakhstan, Luxembourg, Portugal, Russia, South Korea, Spain, United Kingdom and United States. Expansive bills enhance a nation's AI capabilities while restrictive bills impose limitations on AI usage.

Conclusion

AI is undeniably pervasive, with rapid advancements in technology shaping the future of industries worldwide. From the emergence of powerful, multimodal models to the rise of open-source alternatives and retrieval augmented generation (RAG) systems, the state of AI technology is both impressive and evolving. However, despite these technological leaps, many businesses remain in the early stages of AI adoption, driven by fear of missing out rather than a strategic understanding of AI's potential.

To truly harness AI's transformative power, businesses must move beyond superficial implementation. This involves not only embracing the latest technological advancements but also developing a deep, strategic approach to AI integration. This means investing in sustainable, use-case-driven solutions, fostering the necessary skills within the organisation, aligning AI strategies with broader business goals, and navigating the increasingly complex regulatory landscape.

By addressing these challenges and seizing the opportunities presented by current AI technologies, companies can transition AI from a buzzword to a powerful driver of meaningful, lasting change across industries.

References

Ding, Y., Fan, W., Ning, L., Wang, S., Li, H., Yin, D., & Li, Q. (2024). A survey on rag meets llms: Towards retrieval-augmented large language models. . arXiv preprint arXiv:2405.06211.

Epoch AI. (n.d.). Data on Notable AI Model. Retrieved 24 August, 2024, from epochai.org.

Labonne, M. (25 July, 2024). I made the closed-source vs. open-weight models figure for this moment. Retrieved from https://x.com/maximelabonne/status/1816008591934922915

Mearian, L. (22 Aug, 2024). Generative AI is sliding into the 'trough of disillusionment'. Retrieved 26 Aug, 2024, from Computerworld: https://www.computerworld.com/article/3489912/generative-ai-is-sliding-into-the-trough-of-disillusionment.html

NCS. (11 July, 2024). the AI gambit: accelerate with digital resilience. Retrieved from ncs.co: https://ncs.co/en-sg/knowledge-centre/articles/the-ai-gambit/

Precedence Research. (July, 2023). GPU Market Size, Share, and Trends 2024 to 2034. Retrieved 26 Aug, 2024, from precedenceresearch.com: https://www.precedenceresearch.com/graphic-processing-unit-market

Stanford Institute for Human-Centered Artificial Intelligence (HAI). (2023). AI Index Report 2023. Stanford University. Retrieved from https://aiindex.stanford.edu/report/

Tremayne-Pengelly, A. (6 6, 2024). Nvidia's Market Cap Surpasses $3T: Here Are the Largest Buyers of Its A.I. Chips. Retrieved 26 08, 2024, from Observer.com: https://observer.com/2024/06/nvidia-largest-ai-chip-customers/#:~:text=Microsoft%20reportedly%20plans%20to%20amass,during%20the%20same%20time%20frame.

Yeong, Z. (26 August, 2024). Singapore's twin engine approach to AI regulation. Retrieved 26 August, 2024, from Straits Times Interactive: https://www.straitstimes.com/opinion/singapore-s-twin-engine-approach-to-ai-regulation

embarking on the code odyssey: unveiling the power of AI-assisted programming

Leow Wee Sheng, Yap E Fang,
Yong Yam Koon, Vito Chin, Valerie Chan

embarking on the code odyssey: unveiling the power of AI-assisted programming

Leow Wee Sheng, Yap E Fang, Yong Yam Koon, Vito Chin, Valerie Chan with contributions from Brett Spedding (Microsoft)

AI is a strategic imperative

In the ever-evolving landscape of technology, where innovation drives the competitive edge, the role of software development has never been more pivotal. Today, we find ourselves at the intersection of ambition and execution, where harnessing the potential of software innovation is a strategic imperative. The question we now face is not whether to embrace AI, but how to wield it as a catalyst for organisational growth. In this white paper, we shall embark on a journey into the realm of AI-assisted programming and how it revolutionises software development.

The demand for innovation and efficiency in software development has never been greater. Where most industries depend on software as the linchpin of their operations, AI-assisted programming stands as the harbinger of transformation. At its heart, it involves the integration of advanced AI, epitomised by tools like **GitHub Copilot, Amazon Q and NCS CodeNav**, into the software development lifecycle.

In the pages that follow, we will explore the ways in which the triple A benefits of **AI—Assist, Augment, Automate**—can manifest both horizontally and vertically.

We will delve into how we can usher in a new era of software development through three key value drivers, namely development acceleration, code protection, and developer empowerment. You will gain insights into how AI-assisted programming is crucial in helping companies to maintain a competitive edge.

Accelerating software development

Three key value drivers

Traditional software development workflows frequently suffer from inefficiencies and bottlenecks, which have a negative impact on project timeframes and results. These workflows are riddled with repetitive coding tasks that drain valuable time, debugging processes that typically catch errors late in the game, and lengthy development cycles that can stifle adaptability to changing market conditions. This section highlights the pressing need for innovation in the software development landscape and sets the stage for examining how AI-assisted programming offers a game-changing remedy to these age-old problems, ushering in an era of unparalleled productivity.

Benefits of AI-assisted programming

- **Faster coding:** Suggests full functions and code lines, saving time for creative problem-solving.
- **Better code:** Promotes best practices and catches errors, leading to cleaner and more maintainable code.

Quicker time-to-market

AI-assisted programming follows the shift in preference towards agile methodologies – working quickly to make small iterations and get ongoing feedback. Through real-time code suggestions and auto-completion features, these tools simplify the coding process, drastically reducing the need for manual searching and typing. They also assist in the generation of code snippets, the automation of routine coding activities, and the construction of boilerplate code. As such, companies can speed up the software development lifecycle and enable quicker time-to-market in stages such as initiation (code refactoring), requirement & design (user stories generation), construction (code generation), and testing (automated unit test scripts).

Quality improvement

This automation not only accelerates coding but also improves code quality by minimising syntax errors and ensuring best practices are followed. For example, in the case of automated unit test script generation – AI tools like **GitHub Copilot** and **Amazon Q** can analyse the code being developed and provide code suggestions for test cases. It identifies different scenarios and edge cases that should be tested, generating test cases for various input values, corner cases, and boundary conditions. By proactively detecting errors, vulnerabilities, and bugs in real-time, developers reduce the need for debugging efforts later in the development cycle.

GitHub Copilot

Cost reduction

In a **GitHub Copilot** trial conducted to evaluate the impact of AI on developer productivity, it was revealed that participants with access to the AI pair programmer completed the given task **55.8% faster** than the control group[1]. Naturally, the productivity gain would imply a significant amount of cost savings – cost savings in project management that arises from shorter development cycles, and cost reductions associated with bug fixing and maintenance, thereby allowing companies to enjoy significant cost reductions.

Adopting DevSecOps methodology

Securing the digital future

The dependence on software applications has reached unprecedented heights as technology is increasingly incorporated into every aspect of our life. As a result, the effects of software flaws, hacks, and poor-quality code have increased tremendously. These problems not only pose serious risks to the confidentiality and integrity of data but also have far-reaching repercussions for an organisation's standing in the public eye, financial security, and legal standing. To navigate this complex and high-stakes terrain, organisations should start recognising the potential of AI-assisted programming in fundamentally altering how code quality and security are maintained throughout the software development lifecycle. In this section, we will delve into the transformative role of AI in ensuring robust code quality and security, safeguarding against potential vulnerabilities, and fortifying the software that underpins our digital future.

Enhanced code security

AI-assisted programming tools like **GitHub Copilot** offer several advantages that go beyond streamlining the software development lifecycle for agile teams. They also contribute to enhancing code security through proactive vulnerability detection and secure code generation. For example, **GitHub Copilot** employs advanced static code analysis and machine learning algorithms to continuously examine the code being written. It proactively identifies potential vulnerabilities, security loopholes, and coding practices that can expose systems to security hazards. By detecting issues in real-time, it reduces the risk of security and data breaches that can damage the company's reputation and result in legal liabilities.

Adherence to best practices

Tools like **GitHub Copilot** can guide developers to adhere to security coding standards and industry best practices. It offers suggestions for secure coding techniques and guidelines for secure authentication, authorisation, and data protection, helping developers write more secure code from the outset. For example, it can detect and recommend encryption techniques for handling sensitive data in compliance with data protection laws like GDPR or suggest secure authentication mechanisms that align with industry standards. Automated checks for compliance with industry standards and regulations help the company avoid costly penalties and legal complications.

Reduced technical debt

Promoting clean, well-structured code and automated adherence to coding standards help development teams minimise the accumulation of technical debt over time. They prevent shortcuts, hacks, and workarounds that frequently lead to a complicated web of code, making future updates and maintenance increasingly costly and time-consuming. Conversely, AI-driven suggestions guarantee that code is created in a comprehensive and consistent manner, thereby lessening the cost burden brought on by extensive refactoring, codebase rewrites, and extended development schedules, and hence reducing the danger of technical debt accumulation.

Empowering developers for the future

Empowering developers with AI to innovate and inspire

No longer confined to the traditional realms of coding and debugging, developers are emerging as creative problem solvers and innovators, driving the very essence of technological advancement. In this section, we delineate how AI-assisted programming tools like **GitHub Copilot** have accelerated and empowered this change in developers' function, and how these technologies not only speed up development chores but also serve as invaluable mentors, offering on-the-job learning experiences and exposing developers to a breadth of coding styles and techniques.

Upgraded developer skillset

AI-assisted programming tools like **GitHub Copilot** foster on-the-job learning opportunities for developers. These tools, powered by machine learning and code analysis, actively assist developers in real-time, suggesting code snippets and offering insights into best practices and emerging technologies. With recommendations spanning a wide range of programming languages, libraries, and frameworks, developers can diversify their skill set effortlessly. By promoting continuous learning within the development workflow, developers not only gain proficiency in current technologies but also remain primed to embrace emerging trends. This translates to a more innovative and agile organisation, ultimately contributing to the company's competitive edge and long-term success in the ever-evolving digital landscape.

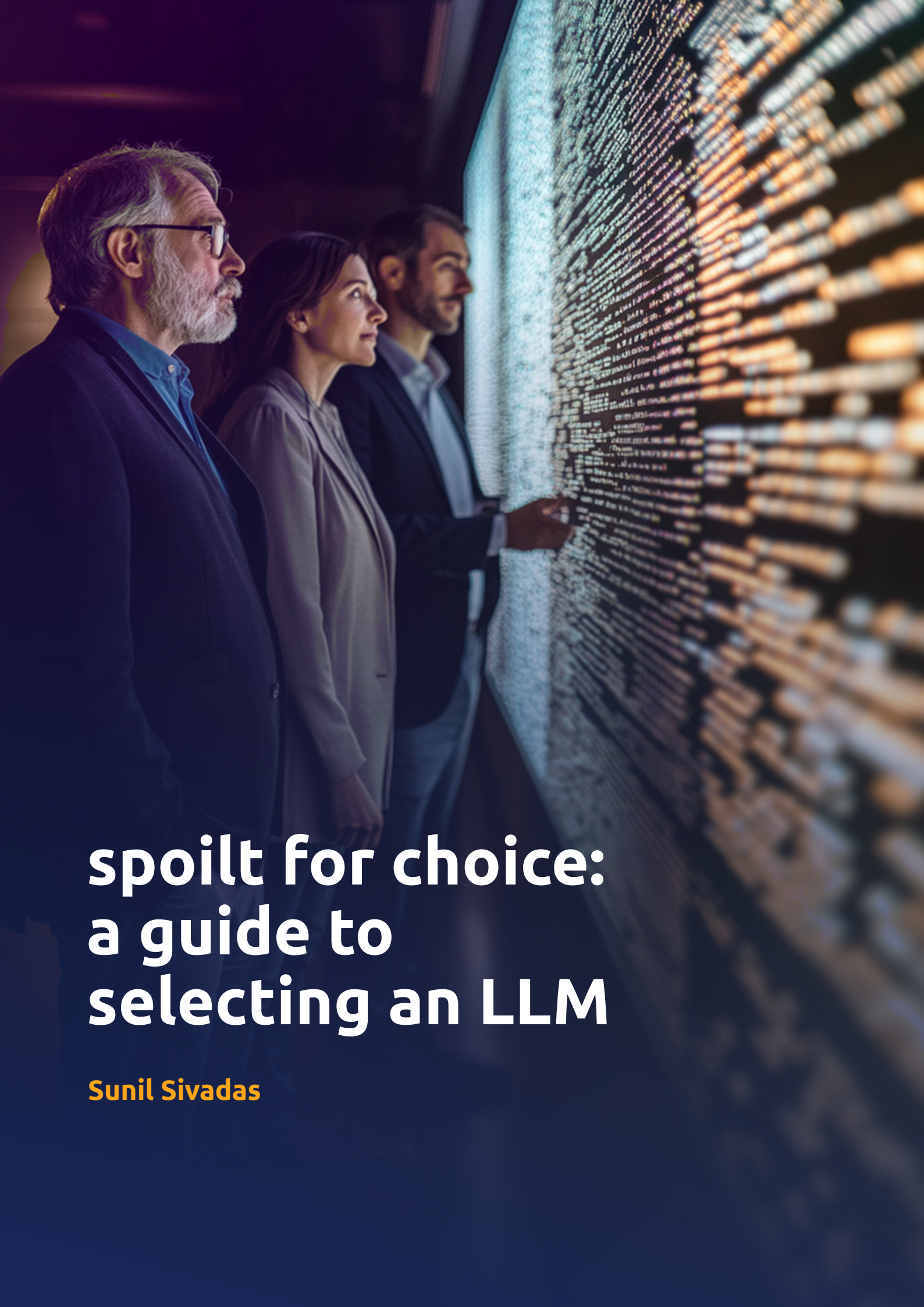
Higher employee retention

Employees who appreciate professional development may find these tools that assist continual learning and skill improvement appealing. In fact, 71% of workers surveyed in a Gallup study felt that job training and development increased their job satisfaction, and 61% indicated that upskilling opportunities are an important reason to stay at their job. Hence, the use of such tools may in turn result in a higher retention rate. What further establishes the long-term retention of employees is the fact that these tools reduce the monotony commonly associated with repetitive and time-consuming coding tasks. Empowering developers of the future to automate routine tasks and instead focus on more meaningful and intellectually stimulating aspects of their work increases job satisfaction. For instance, a survey conducted revealed that 60-75% developers who utilised **GitHub Copilot** felt more fulfilled with their job, felt less frustrated when coding and were able to focus on more satisfying work. Similarly, 73% of developers reported that the tool helped them stay in the flow, and 87% of them could preserve mental effort during repetitive tasks.[1]

Reduced recruitment pressure

Overall, offering cutting-edge tools can make a company more attractive to top tech talents, as skilled developers are more likely to stay with a company that provides them with the tools and resources they need to excel in their roles. The ability to attract and retain these talents is crucial today as programmer scarcity is an increasingly pressing concern for organisations, with the global tech talent shortage expected to reach 85.2 million by 2030[2]. Developer shortage has stalled new projects in recent years, but AI-assisted programming tools like **GitHub Copilot** offer a compelling solution. By upskilling existing talent, Copilot not only mitigates the recruitment pressure but also elevates the company's agility and cost-efficiency. Developers are provided with instant access to a wealth of coding knowledge, gaining assistance to tasks ranging from routine to complex. The acceleration of development cycles also enables current team members to handle more tasks with greater efficiency. This, in turn, reduces the need for extensive external hiring and helps organisations to navigate the developer shortage with confidence.

1 From: https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/
2 https://www.griddynamics.com/global-team-blog/software-developer-shortage-us



spoilt for choice: a guide to selecting an LLM

Sunil Sivadas

spoilt for choice: a guide to selecting an LLM

Sunil Sivadas

Executive summary

Final week of July 2024 marked a pivotal moment for open-source AI development, with two major advancements pushing these models to the forefront of technological innovation and potentially broadening access to advanced AI tools. First, Meta's CEO, Mark Zuckerberg, introduced Llama 3.1, a leading-edge model made freely available to everyone. Just a day later, Mistral, a French AI startup, unveiled Mistral Large 2, a strong competitor to the top AI models. The convergence of this new generation of LLMs with the growing GenAI excitement of business executives since ChatGPT's release signals a paradigm shift in the development of LLMs that will result in: 1) a movement **from consumer to enterprise** applications; 2) a greater **demand for domain-specific models** over general-purpose models; and 3) the emergence of **open-source models with permissive licenses** that could spark new levels of innovation.

As the landscape of available LLMs gets more crowded by the day, companies looking to explore LLMs and GenAI face an increasingly difficult, but fundamental, question: **Which LLM should I choose?** This article proposes a lens through which companies can gain perspective and make an informed decision.

To navigate the sea of LLMs successfully, **awareness of the different characteristics** these models have is crucial. Specifically, it is important to understand if the model is: 1) general-purpose or domain-specific; 2) proprietary or open-source; and 3) deployed as an API, in the cloud, or on-premises?

A deep understanding of one's use case is required to match the right model characteristics to the application requirements. Questions to be answered include: 1) Am I using GenAI to automate a process, or to assist a person's completion of a task; 2) What is the value-at-stake for the target use case; 3) Does my use case require specific industry knowledge; and 4) What are the necessary security requirements for my use case? Understanding the use case will allow companies to identify the suitable characteristics their chosen LLM must offer. However, there are other design considerations that must also be addressed.

On the business side, executives should conduct a **cost-benefit analysis** to understand if the adoption of GenAI will boost or hamper value creation. Should GenAI potentially be a major disruptor to their industry, companies should also **identify the areas of competitive advantage** on which they can capitalise. Lastly, companies must be aware of the reputation of the LLM vendor, and the **level of support** they will receive in running their LLMs.

The technical considerations fall into three distinct categories. First, for data, companies must select their LLMs in accordance with their **data sensitivity** and security requirements, and with data availability, a key enabler for the option of fine-tuning. Next, companies must also be cognisant of the model's **performance on benchmarks**, as well as how **permissive the model license is**, especially so in cases of potential competitive advantage. Finally, deployment options must also be closely considered, and companies need to be aware of the **trade-off between security compliance and infrastructure investment.**

At the current state of LLM development, it is highly unlikely that companies will be able to identify a 'silver bullet' LLM that will provide for all their GenAI needs. But having this informed perspective into evaluating and assessing LLMs will enable companies to explore the available options competently and confidently, and to also make the switch when a golden opportunity presents itself in the future.

Generative AI and current trends

An internal Google memo, that was leaked in May 2023, claimed that Open Source AI would outcompete Google and Open AI, summed up simply with the phrase: "we have no moat"[1].

The release of Meta's LLaMA models gave open-source LLM development impetus, which has led to the creation of models that have shown to be able to 'punch above their weight,' doing much more with much less. Open-source development, far from being a sideshow to the research efforts of Big Tech corporate labs, has begun to pave a crucial, divergent path in the space of LLM creation and development.

That divergent development, compounded with the growing excitement of business executives for Generative AI (GenAI) following the release of ChatGPT, has resulted in three key shifts in the GenAI and LLM market that we believe will have a profound impact on businesses. Namely, (1) increased business adoption of GenAI; (2) divergent trends in LLM development (general purpose vs. task-oriented); and (3) a notable change to permissive open-source LLM licenses.

Business executives are now flocking to implement some form of GenAI in their companies in the hope that it can help them to improve their business outcomes. Seeing a growing market of unmet demand, some GenAI startups have pivoted their solutions from the consumer to the enterprise space. Tome, a startup that offers an AI-powered tool designed to help users create dynamic, visually compelling narratives without the need for traditional slide decks, shifted its focus from consumer products to more revenue-generating enterprise space[2].

As industries continue to bring their GenAI use cases forward, there is a divergent trend in how LLMs are being developed. While LLM development has generally been dominated by hyperscaler LLMs, there is also a wave of industry-specific LLMs that are being trained with industry knowledge to perform better on high value use cases. Open-weight models such as Codestral, code-specific LLM released by Mistral, are challenging products such as GitHub CoPilot by driving the adoption of self-hosted coding assistant[3].

As open-source communities find their footing to compete with Big Tech resources, we observe that more permissive open-source LLM licenses are gaining traction.

Rapid proliferation of LLMs

As it stands, the landscape of LLMs is crowded, and will get even more so week by week. Companies intending to explore the capabilities of LLMs are faced with a simple but fundamental question: Which one? The rapid proliferation of LLMs in the market, because of GenAI hype, is leading to 'choice overload.' We believe that deeper clarity in how to understand the characteristics of these new models will help companies to make more informed decisions.

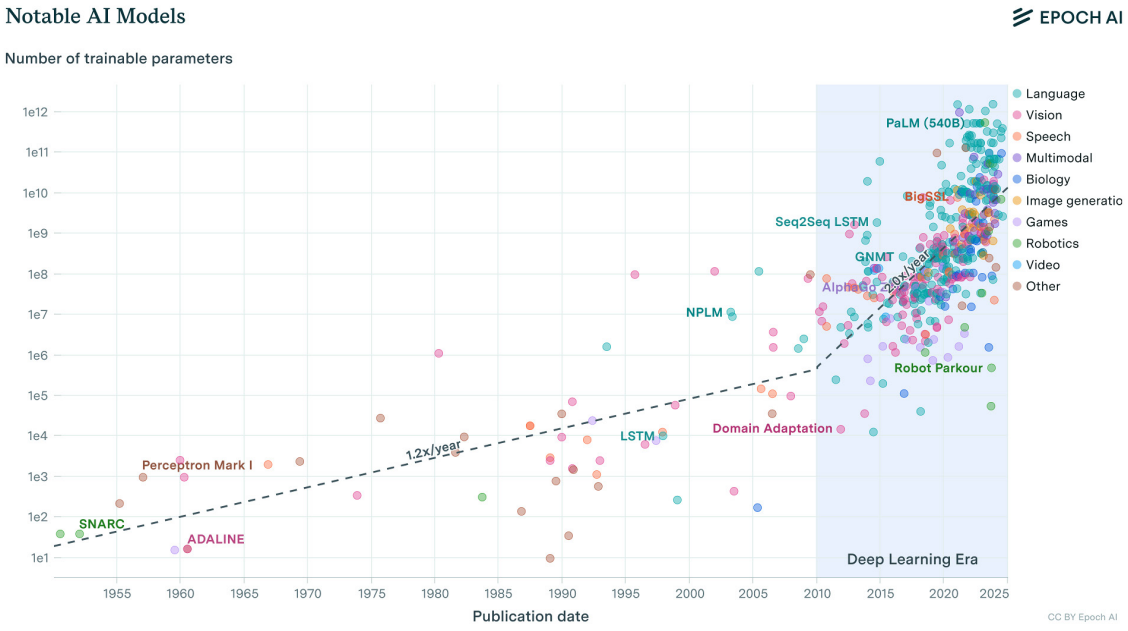


Figure 1. Prominent AI models for different modalities and their growth in number of weights or trainable parameters. Source: Epoch AI 15 August 2024.

These considerations span across the model type, use case, business design and technical design, and have been detailed out in the next few sections.

Figure 2. Overview of framework for selecting an LLM

Model characteristics

General-purpose vs. domain-specific

Almost all the most noteworthy LLMs, including GPT-4, Claude, Gemini, and most recently, Llama 3.1, are general-purpose models. Having been trained on material scraped from the internet, these models can perform a wide variety of tasks, on an endless range of topics, and are like an AI-powered encyclopaedia virtual assistant. A user would just need to be able to prompt the model properly to get the desired output and information. It is the seemingly limitless capabilities of these models that have driven much of the excitement surrounding LLMs.

However, there are other LLMs trained with a specific context or industry in mind. Some examples include Med-PaLM for medical knowledge and Codestral for coding applications. These models are trained to be more accurate and reliable for a chosen field or use case and are more like subject matter experts for that industry.

Proprietary vs. open source

Many of the early LLMs were also proprietary, with the companies doing research on them mainly publishing research papers, and only providing limited access on request to selected users. This behaviour has continued until today, with many noteworthy models not fully released to the public with access restricted through an API. For the most part, interaction with these models is done via prompt engineering, but there are providers who also allow for some customisation through fine-tuning.

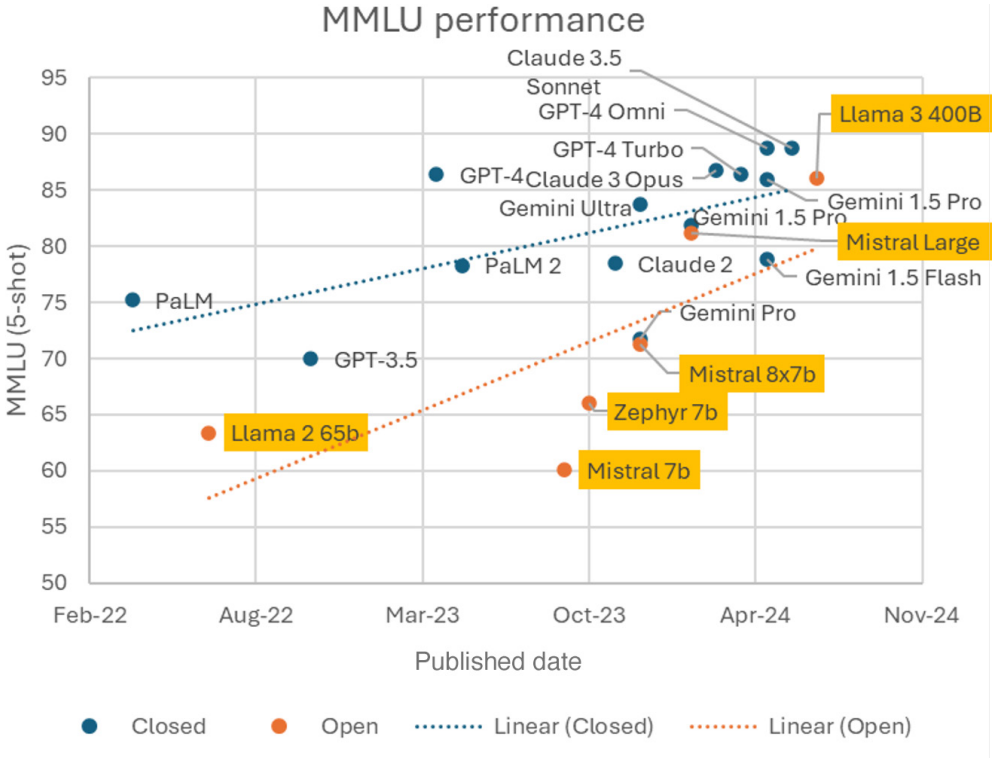


Figure 3. Open-weight models closing the gap with closed models. Source: Maxime Labonne, Twitter, 25 July 2024.

In recent times, however, open-source communities have begun to catch and have developed high-performing models that can compete with the performance of proprietary LLMs. Offering varying degrees of license permissiveness, the creators of these models allow anyone to be able to download the actual model file, and experiment with it, enabling greater flexibility and customisability, not just in terms of the model outputs, but also in deployment.

API vs. cloud vs. on-premises hosting

In terms of deployment, LLMs have been deployed in many ways. For proprietary models like GPT-4, users do not have direct access to the models themselves but interact with them through an API using prompt engineering. Users can pay for API usage through a subscription or on a pay-per-use basis, where the user pays for tokens which are consumed when prompting the model and when generating the content requested. One concern over the use of an API is the lack of trust regarding the security of the application in handling the data being fed into it.

Other providers can help users to host their custom models on a private cloud. Under this option, users are generally able to fine-tune their models as well as ensure data is not being mismanaged, all while not having to invest in expensive compute infrastructure. While this helps to address the concern of data privacy and customisation, users are expected to pay for all the services they require, which includes fine-tuning costs on top of the inferencing costs. In addition, a custom model that has been fine-tuned with the user's data is not considered the intellectual property of the user themselves.

If the data required to prompt the model is extremely sensitive, or if there is a critical need to protect a user's IP to maintain competitive advantage, users can opt to host their LLM on-premises. While this option gives the user a greater level of flexibility and control over the LLM, the data, and the resulting IP, it comes with its own set of challenges including the need to invest is expensive compute infrastructure, substantial engineering, and the recruitment or training of talent to ensure the LLM's capability to affect business outcomes. To address this, infrastructure technology providers are building tools to reduce the complexity of solutions and to increase developer productivity. In March 2024, NVIDIA announced NIM (NVIDIA Inference Microservices), a set of easy-to-use microservices designed to accelerate the deployment of generative AI models across the cloud, data centres and workstations[4].

Use case requirements

To effectively identify which LLM is most suitable, a good understanding of the requirements of the use case is crucial. While each use case will be unique, there are helpful ways to examine use cases to gain clarity about pain points and how the LLM can be integrated to address those points.

Domain knowledge

There are potential use cases that will require very specific domain knowledge such as those in the medical, legal, and coding industries, to name a few. The knowledge and training requirements of the model will limit the suitable LLMs available, and it is also possible that no LLM has yet been trained on the desired knowledge base. If a readily available domain specific LLM cannot be found, other options, such as fine-tuning a base LLM, will need to be explored.

GenAI purpose (automate vs. assist)

One key consideration for integrating GenAI is understanding the degree of automation required for the particular use case. If it can be expected that the model will be left to automate some processes on its own, the tolerance for error will be low. On the other hand, if the GenAI solution is there to assist users in completing their tasks, having that human-in-the-loop to curate and screen the model's outputs will mean that the tolerance for error will be much higher. Knowing how GenAI fits into the workflow of operations will help users understand what level of performance is expected from the LLM and if additional measures like fine-tuning are necessary.

Value-at-stake

A closely related consideration is the value-at-stake of the use case itself. If the use case is one where there is high value-at-stake (i.e., large contract amount, medical applications), the error tolerance will be very low, so the output of the LLM used must be very reliable so as not to jeopardise the solution. But if the use case's value-at-stake is much lower, the inherent variance of model output can be acceptable and an LLM that performs just 'well enough' can be tolerated.

Security

Some use cases may deal with highly sensitive matters, such as personal or high value information, or possibly national security interests, in the case of government agencies. In these situations, the security of the data used for prompting and fine-tuning, as well as the model outputs themselves, must be treated with the utmost vigilance. The security requirements of the use case will affect the chosen deployment method, which then affects the range of suitable LLMs available to the user.

With a deeper understanding of the use case, selecting a suitable LLM with characteristics that match the requirements becomes a simpler task. Use cases that require a high reliability of model outputs will require an LLM that has been proven to perform at a high level. If none are available, selecting an LLM that supports fine-tuning becomes crucial. Use cases that require specific domain knowledge will be best served either by existing LLMs that are already trained with the necessary data, or by LLMs that can be fine-tuned to meet that requirement. Use cases that have high security requirements will necessitate either a private cloud environment or even on-premises hosting to ensure the necessary security standards.

Additionally, there are other design considerations that companies must take into account in evaluating, and ultimately selecting, an LLM.

Business considerations

Fundamentally, the use of the LLM should help an organisation to better achieve its business outcomes. Implementing GenAI, without a clear vision, will create distractions instead of creating value. One key calculation for companies looking to implement GenAI is a cost-benefit analysis. Balancing the capital expenditures (compute resources, GPUs) and operating expenses (licensing fees, salaries) of investing in LLMs with the value they create in terms of cost savings, productivity, or new revenue is critical, and companies should begin with the end in mind, and aim towards a positive return on investment. An interesting trend is LLM providers are passing on the gains from optimised implementations and decreasing cost of hardware to customers by charging less even as model performance rises.

A significant trend to note is the reduction in prices by large language model providers for their customers. This is due to enhanced efficiency in implementation and decreasing hardware costs, even as the performance of these models continues to improve. The cost to achieve an Elo rating of 1250 has dramatically dropped by a factor of 100, from roughly $30/million tokens for Claude 3 Opus in March 2024 to $0.30/million tokens for gpt-4o-mini in July 2024[1]. Following a recent price cut by OpenAI, GPT-4o tokens are now priced at $4 per million tokens, based on a blended rate that assumes 80% input and 20% output tokens. When GPT-4 was initially released in March 2023, it cost $36 per million tokens. This price reduction over a span of 17 months equates to an approximate 79% annual drop in price.

Another consideration is the degree to which implementing GenAI will generate a competitive advantage for the business. If the integration of GenAI can potentially be a key differentiator, or if GenAI could be disruptive to one's own industry, companies would be well-advised to explore LLMs closely to ride the wave of disruption, or risk sinking instead. Being cognisant of the shifting trends will help industry executives anticipate the opportunities and threats that GenAI will pose to their businesses.

Finally, executives should also consider the reputations of the vendors providing LLMs. Working with LLMs can be a tricky process, and the level of support provided by the vendor is crucial to a successful exploration. For open-source LLMs, having clear documentation and community support will go a long way toward simplifying the installation and deployment processes.

Technical considerations

There are numerous technical considerations that companies must also be aware of in selecting an LLM, among which data management and security are key. Not all LLMs can be integrated into a secure environment, so selecting an LLM that can be hosted securely will be crucial for companies that possess a lot of sensitive data. Data availability is another important consideration when selecting an LLM. A company that has large amounts of high-quality data can leverage their data resources to engineer better prompts, or even perform fine-tuning, and should select an LLM that can support this capability.

For use cases that require higher model output reliability, it may be better to explore a customisable model that can be fine-tuned for a specified use case using proprietary data. If data is not as readily available, companies should target LLMs that have performed well against common benchmarks. In the case of open-source models, if the resulting solution will potentially become an important source of competitive advantage, LLMs with more permissive licenses should be considered ahead of those that are less permissive, so that the IP can be protected.

Another important factor is the benchmarking performance of the LLMs. There are currently different benchmarks available that are used to evaluate and rank the performance of LLMs on various applications and use cases and are equated to standardised testing for language models. Leaderboards, such as those prepared by LMSYS5 and Hugging Face6, provide visibility by creating a snapshot of the benchmarking performance of various LLMs in comparison to each other. Companies must also note that while benchmarks may be helpful tools in comparing LLM performance, they do not have the inherent ability to evaluate the quality of the content generated by the models in real world situations which is a difficult task and remains an open research challenge.

If ensuring data privacy or IP protection is critical, companies should consider LLMs that can be deployed in secure environments. Additional deployment considerations are the necessary operational requirements of the model, such as the compute infrastructure and engineering effort required, and the talent needed to execute the various LLM operations. Companies must be aware of the resources currently available to them and decide if they would prefer to invest in more resources to use a larger, more complex model, or to find a model that can be deployed given their existing levels of resources.

At the current state of LLM development, it is highly unlikely that companies will be able to identify an LLM that is a potential "silver bullet" for all their GenAI needs. Complex use cases may even require the deployment of a portfolio of LLMs, used in concert, to produce the desired output, provided that available resources allow for the use of multiple models.

1 Smol.AI, 8 Aug 2024

Conclusion

The framework proposed should help businesses to objectively assess available LLMs and to decide on one (or a few) to explore and, we hope, make the process less daunting. Selecting an LLM is just one of the initial steps and should not become a roadblock in the exciting process of GenAI exploration.

As this space continues to evolve rapidly, having an informed perspective on one's use case and the available LLMs will be crucial for companies in distinguishing between noise and necessity. Most of the new releases should not delay or derail a company's planned exploration of GenAI using a selected LLM, but keeping a keen eye on ongoing LLM releases and developments will help companies to identify important switching opportunities as and when they present themselves.

Generative AI solutions for Automated Speech Recognition (ASR)

Chong Yang Ng, Cliff Tan

Generative AI solutions for Automated Speech Recognition (ASR)

Beyond words: delving into deeper conversational insights

Chong Yang Ng, Cliff Tan

Executive summary

Automated Speech Recognition (ASR) is a common tool used by many businesses to ensure that customer conversations are captured and retained for customer service, legal, and compliance purposes. But linguistic factors can create challenges to ensuring the accurate transcription of conversations, particularly in regions as linguistically diverse as Southeast Asia.

Based on our experience, most AI solutions have challenges in interpreting our local accented speech which often contains a mix of borrowed words from different languages and dialects, idiosyncratic syntax and grammar as well as references to local places.

In this article, we explore the immense potential of ASR to revolutionise business operations, empower growth, and drive unparalleled conversational understanding. Additionally, we will address the current challenges in ASR capabilities, especially in an ASEAN context.

Challenges with ASR capability in an ASEAN context

Automated Speech Recognition (ASR) is a technology that converts spoken language into written text. While ASR technology has been around for some time, several challenges persist, particularly in the ASEAN context:

- Diverse linguistic landscape: The ASEAN region is rich in cultural and linguistic diversity. With multiple languages, accents, and dialects spoken across countries, developing an ASR system that accurately recognises and transcribes all variations presents a considerable challenge.
- Code-switching and creole languages: Many ASEAN countries have a culture of code-switching, where individuals switch between languages in a single conversation. Additionally, the presence of creole languages further complicates ASR accuracy, as these languages often lack standardised grammatical rules and structures.
- Acoustic environment variability: The ASEAN region encompasses diverse acoustic environments, including bustling urban centres and remote rural areas. Background noise, varying speech volumes, and microphone quality can impact ASR performance, requiring robust adaptation techniques.
- Low-resource languages: Some languages in the ASEAN region may have limited data available for training ASR models. The scarcity of training data for low-resource languages hinders the development of highly accurate and robust ASR systems for these languages.
- Accents and pronunciation: Different accents within a single language can lead to misinterpretations by ASR systems. Pronunciation variations can also pose challenges, especially for ASR models designed primarily for standardised speech.

Understanding NCS' ASR capability and Generative AI

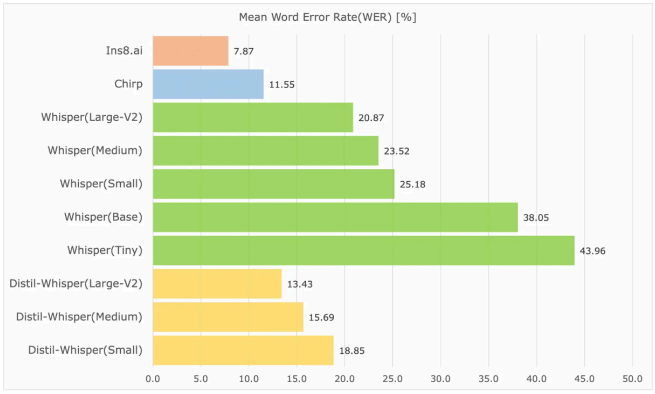
NCS' very own ASR product, Ins8.ai, is an advanced hyperlocal large vocabulary continuous speech recognition (LVCSR) product developed by the NCS Ins8.ai product team and NCS NEXT Gen Tech teams. It was trained using local call centre data on multiple ASEAN languages and excels in handling accents, dialects, and creole languages, making it ideal for Asia. The core technology is a deep neural network architected and fine-tuned by the Ins8.ai team to ensure exceptional accuracy for spoken creole languages in Asia, such as Singlish. Ins8.ai overcomes the barriers from accents and extended vocabularies that can be challenging for traditional ASR technologies.

Comparison with open-source ASRs

In our comprehensive comparison between the NCS Ins8.ai product and Open AI's Whisper (ASR system trained on 680,000 hours of audio datasets collected from the web, two-thirds of which are in English), the results clearly demonstrated Ins8.ai's superiority. Ins8.ai outperformed, producing the lowest word error rate (WER) despite maintaining a small model size (just one-sixth that of Open AI's Whisper medium model) for typical call centre conversations in Singlish while running at a much higher speed of transcription (Figure 1).

Speech-to-text accuracies

From our testing against Singaporean accented English audio, ins8.ai achieved the best accuracy when compared to other models. It scored the lowest WE values despite maintaining a small model size. Below, we provide further details such WER, processing steps, models used, limitations, etc.



Organisation	Model	Parameters(M)
NCS Pte Ltd	Ins8.ai	120
Google	Chirp	2,000
OpenAI	Large-V2	1,550
OpenAI	Medium	769
OpenAI	Small	244
OpenAI	Base	74
OpenAI	Tiny	39
HuggingFace	Distil-Large-V2	756
HuggingFace	Distil-Medium	394
HuggingFace	Distil-Small	166

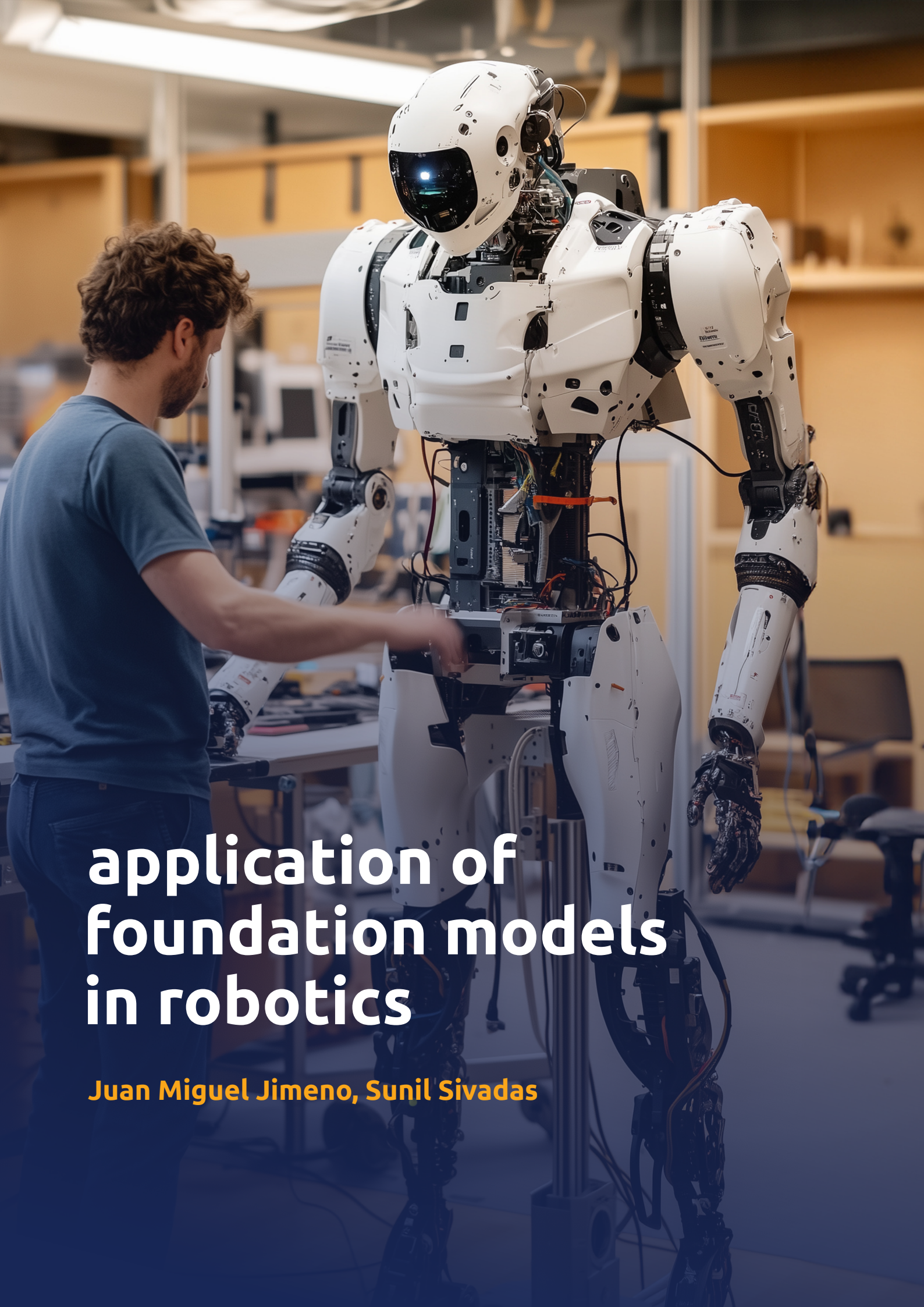
Figure 1. Comparison of NCS Ins8.ai and OpenAI Whisper – Transcription of 150 mins of speech

In a recent implementation with a 250-strong call centre in Singapore, our ASR was applied successfully to live speech-to-text transcription for Singlish, with an accuracy rate of 95% with the ability to interpret Singlish terms like "shiok", "can-can" and "Wah-low" as well as local places in Singapore. This is in line with the statistics above and compares favourably to out-of-the-box STT which generally has an accuracy rate of 70-80% for Singlish.

Beyond call centres, this capability can also be applied to other use cases, such as:

- Financial services – Transcribe financial consultations (between Banking Relationship Managers or Insurance Financial Advisors and their customers), conversations involving the execution of transactions
- Healthcare – Automate capture of interactions between doctors and patients, or even amongst multiple doctors treating a patient
- Investigations – Transcribe conversations between investigation officers and suspects, as well as calls made by the suspect to accomplices.

Our dedication to developing cutting-edge technology has led to Ins8.ai's unparalleled performance, making it the clear choice for anyone seeking industry-leading automatic speech recognition capabilities.



application of foundation models in robotics

Juan Miguel Jimeno, Sunil Sivadas

application of foundation models in robotics

Juan Miguel Jimeno, Sunil Sivadas

Introduction

The need for adaptive solutions that can handle dynamic environments has led to a paradigm shift towards robotic systems with increased autonomy and the capability to thrive in unpredictable settings. This shift is most evident in the rise in popularity of autonomous mobile robots (AMRs) over automated guided vehicles (AGVs) that normally operate on fixed routes and have difficulty adjusting to changing environments. According to The Robot Report[1], it is forecasted that 50% of all mobile robot shipments will be AMRs and less than 25% will be AGVs by 2025.

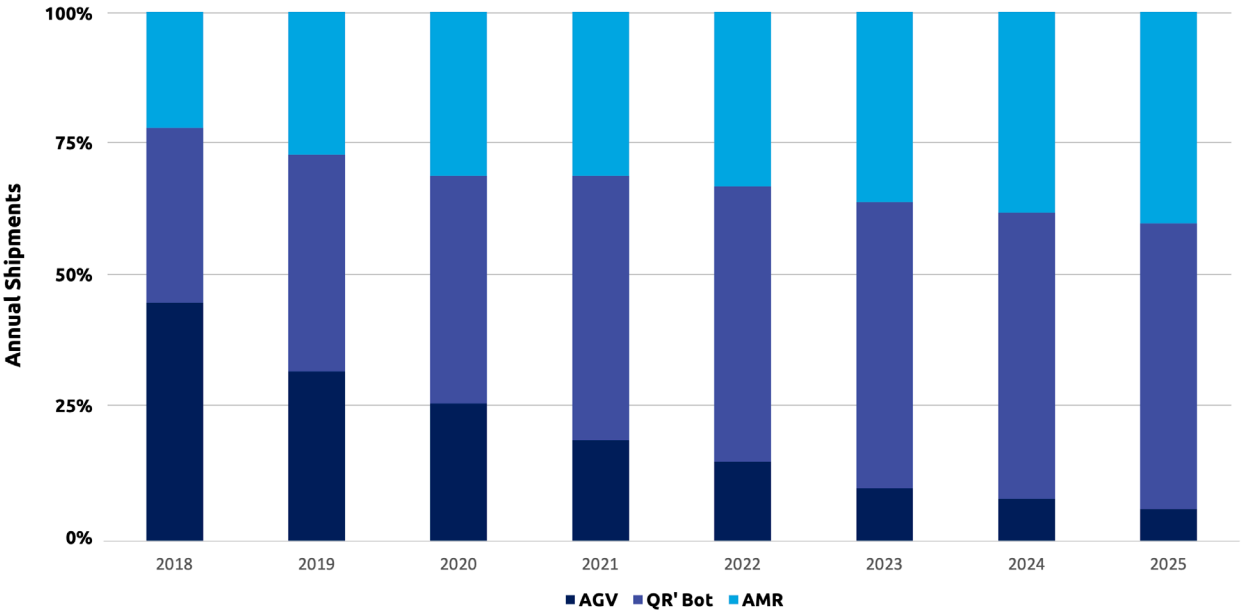


Figure 1. Forecast Share Shipments of AGVs vs AMRs. Source: https://www.therobotreport.com/mobile-robots-rapidly-mainstreaming-by-2025-agvs-and-amrs-could-be-deployed-in-53k-facilities/

Despite this push towards autonomy, robotics task planning is still dominated by classical methods that are based on rigid, hard-coded instructions. However, there is a growing recognition of the limitations inherent in this method given the demands of modern use cases.

Narrow AI

Narrow deep learning model use in robotics applications has gained momentum across applications for use cases ranging from perception tasks that use object detection/segmentation models for finding objects of interest, to training specific robot skills using deep reinforcement learning techniques. In the context of robot task planning involving human-robot interactions, the most common approach typically employs hand-engineered rules paired with deep learning models.

These models are trained to find contextual relationships between the user's input (the task) and pre-determined keywords (commonly referred to as 'intents') that act as triggers for numerous pre-programmed sequences. While offering determinism to the system (as all behaviours are known to the designer beforehand), this approach often struggles to generalise when system alterations to location, time, or logistics occur, and with the introduction of new robot functions. Additionally, scaling proves challenging, particularly with multi-step instructions, due to the intricacies of crafting nested rules for all possible permutations.

Foundation models

Foundation models, on the other hand, are pre-trained on massive internet-scale data and can be adapted to suit diverse tasks. These models have demonstrated notable advancements such as OpenAI's Dall-E2, an AI model that can generate realistic images and art, and GPT-33, the language model that serves as the foundation for ChatGPT. The advent of foundation models has created new options, as researchers start to explore their use in equipping robots with common sense knowledge and the ability to interpret human instructions in a more intuitive manner. A simple task such as "Get me a can of soda" may appear trivial to humans but has proven to be challenging for robots thus far. This challenge arises from the need for robots to not only execute complex algorithms to complete the task, but to also comprehend human instructions and translate them into actionable steps that fall within the robot's capabilities.

In this article, we will discuss recent advancements in the application of foundation models and how these models could transform the way robots are programmed, moving towards a new paradigm of vision and text-based training.

GPT for robotics

In the *SayCan* paper published by GoogleAI4 last year, researchers used a large language model (LLM) to interpret user instructions and rank the likelihood of success for each available skill that the robot would eventually use to accomplish a given task. Their proposed decision-making method achieved an 84% success rate in overall planning.

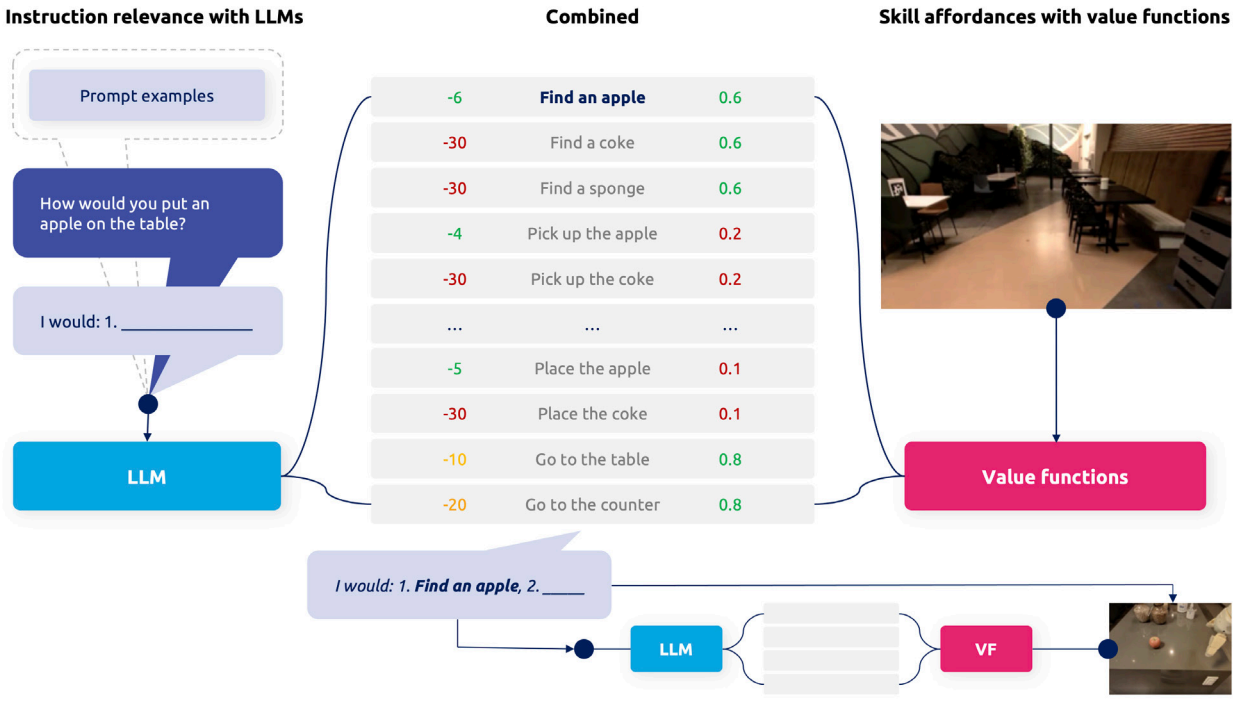


Figure 2. Images of SayCan scoring relevant functions needed to accomplish the task. Source: https://sites.research.google/palm-saycan

The release and use of open-source foundation models gave robots advantages beyond task planning, and could empower them to seamlessly engage in natural language understanding, generation, and interaction, broadening the scope for communication and collaboration on various general-purpose tasks. Peter Chen of Covariant AI referred to the application of principles similar to ChatGPT as "GPT for robotics"[5] signalling the upcoming frontier for foundation models. These models, designed to tackle broader tasks instead of domain-specific problems, are trained on extensive datasets and provide robots with intelligence that can be generalisable across multiple tasks.

Beyond text inputs, robots could also benefit from multi-modal foundation models that not only facilitate word comprehension, but also combine various senses as inputs to enhance robotic understanding of an environment. In September of 2023, OpenAI released an update for ChatGPT that allows the model to "see, hear and speak." This new feature set allowed ChatGPT users to upload photos and ask the model questions that are related to the image.

Vision-language-action model

Recently, researchers have been experimenting with the same multi-modal approach (particularly using images) to aid robots with visual perspective for situational awareness and to learn new generalisable skills. In a recent TechCrunch interview6, Ken Goldberg, a professor at UC Berkeley and the Chief Scientist of the robotics parcel startup, Ambidextrous, remarked that "2023 will be remembered as the year when Generative AI transformed Robotics" as roboticists discover that large vision-language-action (VLA) models can be trained to allow robots to see and control their motions.

Google Deepmind researchers recently trained RT-27, a VLA model, with web-scale data and a robot's past experience to predict the robot's actions and represented those actions as strings. The key objective of this work was to develop an AI model that can learn how to map what it sees (robot's camera view) into robot-specific actions. Compared to SayCan, which acts as a decision maker by ranking the most suitable action based on contextual reasoning and feasibility, RT-2 directly controls the robot based on its visual and language interpretations.

The Google Deepmind team's work could unlock reasoning abilities, that had previously been challenging or nonexistent, for robots. For instance, in 'worst-case' situations when a robot encounters navigation errors, or gets stuck after avoiding an obstacle, traditional methods often struggle to recover, especially on edge cases that were not accounted for during development. In fact, there are companies that provide human-in-the-loop (HITL) services to remotely enable robots to resume operating in these situations, in order to increase the robot's operating time to meet service level agreements (SLAs). With VLAs, operators could teach robots safe and effective recovery behaviours through hours of tele-operation videos and past HITL data.

Tasks such as navigating through an HBD (Housing Development Board) void deck to locate lifts, which usually requires an intuitive understanding based on past experience, can also be made easier for robots. Robots could be trained to make decisions about where to turn, and to choose paths that have a higher probability of leading to a lift lobby, based on visual cues. This skill would be extremely useful for last-mile delivery robots when transiting from outdoor environments to indoor HDB units.

Research performed by a team at UC Berkeley ("Navigation with large language models")[8] demonstrated that robots can leverage the semantic information in large language models (LLMs) for navigating unfamiliar environments using the images from the robot's camera. The LLM's reasoning skills were used as the decision-making mechanism for determining the direction the robot should take to reach the desired goal or location, or to find the object of interest, based on what the robot perceived.

Figure 3. "Navigation with large language models"
The LLM directing the robot towards the refrigerator and microwave as it has higher likelihood of finding the gas stove nearby.
Source: https://sites.google.com/view/lfg-nav/

Simulation for data generation

To meet the huge data demands for training AI models, Generative AI could serve a unique function by generating data and simulating robot experiences. Dhruv Batra, the Research Director at Meta's FAIR (Fundamental AI Research) lab, highlighted a compelling application of Generative AI10 to produce 2D images, videos, and 3D scenes. This could be done using simulators to accelerate scene building and asset generation (quoting one of Deepu Talla's use-cases11 on how Generative AI could contribute to the future of robotics).

A great practical example of this is the Isaac Simulator plugin9, developed by Nvidia engineers, that enables users to create digital twin environments with a simple prompt that describes the desired 3D scene and the assets of interest. The plugin leverages ChatGPT to identify suitable furniture that is available in the simulator's database and to arrange that furniture in a way that fulfills the desired simulation setup.

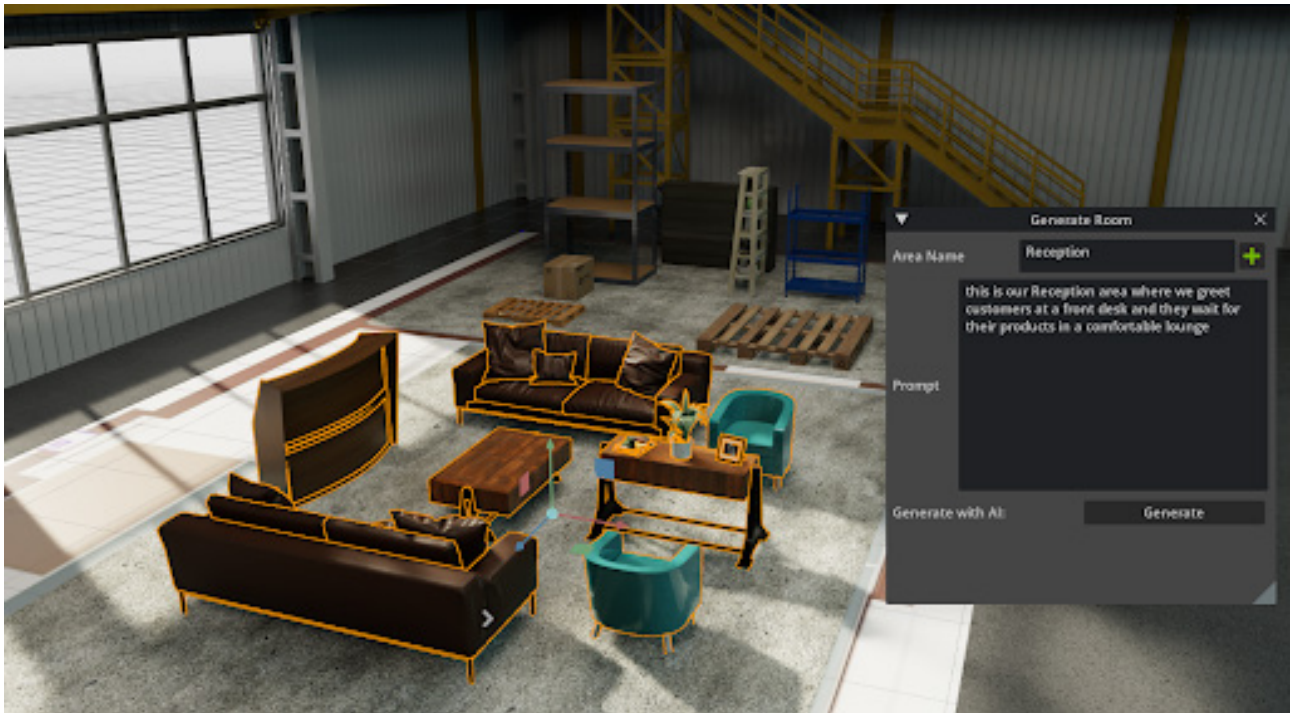


Figure 4. Furniture that was selected and arranged by NVIDIA's LLM-based Isaac Simulator plugin to create a simulated environment of a living room. Source: https://github.com/NVIDIA-Omniverse/kit-extension-sample-airoomgenerator

Humanoid race

For years, Boston Dynamics' Atlas has been at the forefront of humanoid robotics, showcasing impressive feats of agility and balance. However, the field is rapidly evolving with new companies like 1X, Figure, and Tesla's Optimus project joining the race to develop advanced humanoid robots. Having the shared goal of building general-purpose robots, these companies have been pushing the boundaries of what's possible to create machines that can seamlessly integrate into human environments and perform a wide variety of tasks. Figure13 and 1X14 have recently partnered with OpenAI, underscoring the growing importance of foundation models in accelerating humanoid development.

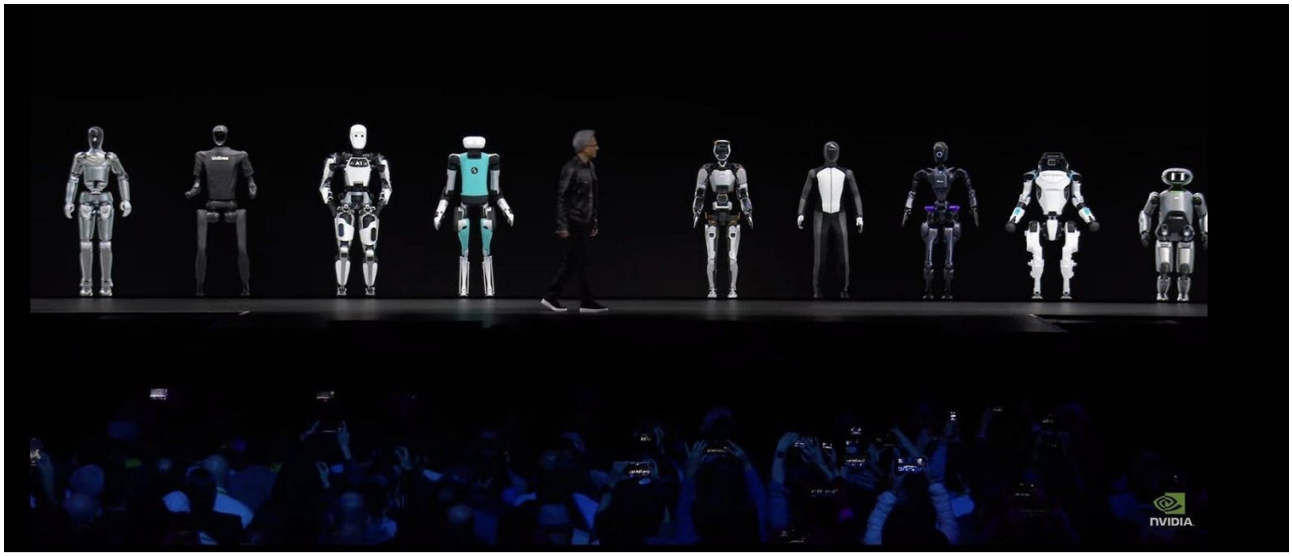


Figure 5. NVIDIA GTC GR00T release. Source: https://www.1x.tech/discover/1x-humanoid-robot-neo-featured-in-nvidia-gtc-keynote

Adding to this momentum, NVIDIA announced Project GR00T15 during GTC in March 2024. This initiative aims to create a general-purpose foundation model for humanoid robots, capable of processing multi-modal instructions and previous interactions. Leveraging NVIDIA's advanced cloud-based GPU infrastructure optimised to run computer graphics suitable for simulation and training, Project GR00T enables robots to perform diverse tasks through both high-level reasoning and low-level motion control. The model's ability to integrate self-observation with multi-modal learning techniques allows humanoid robots to react dynamically to their environment, potentially accelerating skill acquisition and development processes.

The use of foundation models in humanoid robotics presents major advantages as these AI models have been trained on vast amounts of internet data, allowing them to map seen videos and observations into useful robot actuations to perform tasks. By leveraging the diverse and extensive data available on the internet, foundation models can "understand" a wide range of scenarios, objects, and actions. When applied to humanoid robots, this understanding translates into more adaptable and capable machines. For instance, a robot equipped with a foundation model might observe a human performing a new task – like folding a particular type of clothing or operating an unfamiliar device – and be able to replicate that action without explicit programming for that specific task.

Moreover, this ability to map diverse data to physical actions enhances the robot's interaction with its environment and with humans. It can lead to more intuitive human-robot communication, as the robot can better interpret natural language instructions or even non-verbal cues, translating them into appropriate actions. This paves the way for humanoid robots that can assist in various settings—from homes to hospitals to factories—with greater flexibility and understanding of context.

Multi-robot coordination

Early in 2024, Google Deepmind released AutoRT[12], a system that combines large foundation models like LLMs and VLMs with robot control models like RT-2. This integration allows the system to deploy robots equipped with cameras and manipulators for various tasks in diverse environments. The VLM interprets the robot's surroundings and identifies objects, while the LLM suggests creative tasks for the robot to perform. AutoRT can control multiple robots simultaneously, orchestrating them safely in real-world settings. The new system was mainly designed to collect data for robot training and worked by assessing the robot's performance on executed tasks against previously recorded videos. Over seven months of extensive evaluations, the system successfully managed up to 20 robots at once and a total of 52 unique robots. Deepmind conducted 77,000 trials across 6,650 different tasks in office buildings.

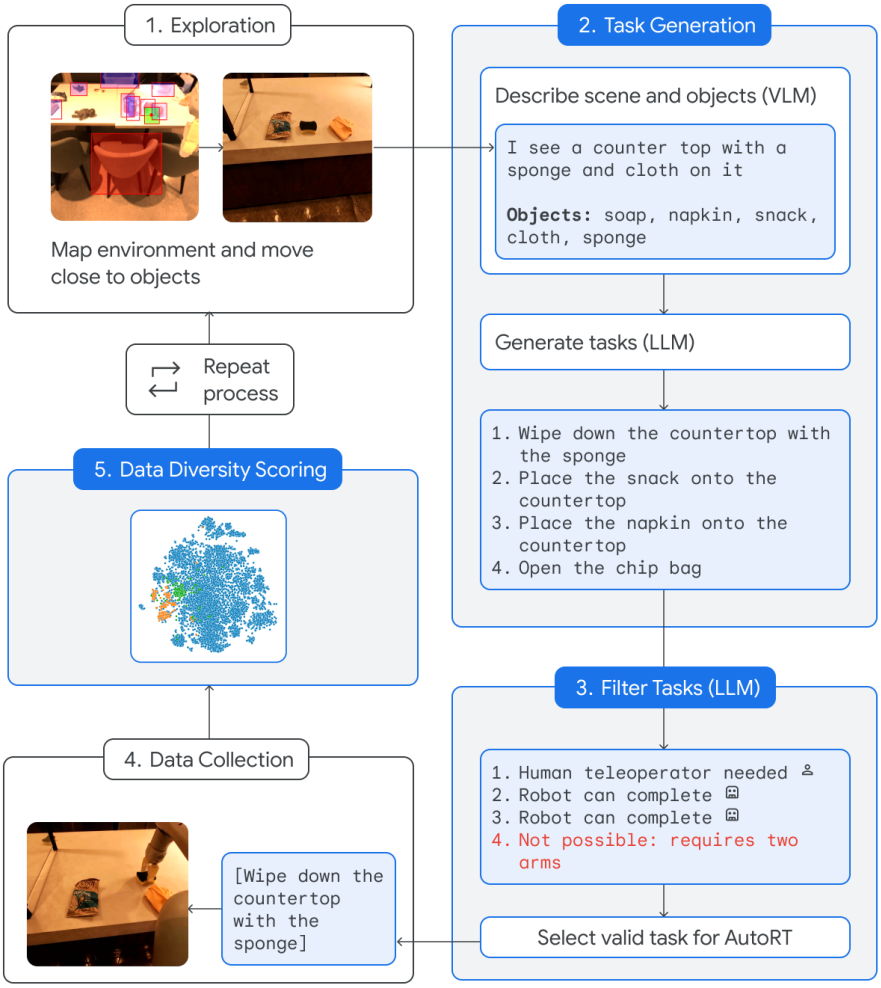


Figure 6. AutoRT's 5-step robot episode data collection process. Source: https://auto-rt.github.io

Although AutoRT is a data collection system, it demonstrates the potential of foundation models to coordinate robot fleets and offers a powerful tool for efficiently managing robots that can adapt in dynamic environments. The utilisation of this system could greatly benefit platforms like Robotmanager by allowing users the flexibility to transition from manual task assignments to an automated, intelligence-driven approach. Such an approach promises to help organisations optimise resources, streamline operations, and reduce the amount of human involvement currently required to control and manage robot fleets.

Summary

The application of Generative AI in robotics holds immense potential for revolutionising the field by addressing key current challenges in human-robot interaction, task planning, multi-robot coordination, and data augmentation. The use of foundation models, such as vision-language models (VLMs) for enhanced perception, and large language models (LLMs) for task planning, will enable robots to better comprehend human instructions and to generate diverse plans for accomplishing user tasks. The multi-modal capabilities of VLMs, in particular, can play a pivotal role in augmenting robot perception, allowing for a more comprehensive understanding of the environment. Additionally, the integration of systems like AutoRT will facilitate effective multi-robot coordination and enhance collaborative efforts in complex scenarios.

References

1. Ash Sharma, Mobile Robots Rapidly Mainstreaming by 2025, AGVs and MARS Deployed in 53K Facilities. https://www.therobotreport.com/mobile-robots-rapidly-mainstreaming-by-2025-agvs-and-amrs-could-be-deployed-in-53k-facilities

2. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever. Zero-Shot Text-to-Image Generation. https://arxiv.org/abs/2102.12092

3. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. Language models are few-shot learners. https://arxiv.org/abs/2005.14165

4. Michael Ahn , Anthony Brohan , Noah Brown , Yevgen Chebotar , Omar Cortes , Byron David , Chelsea Finn , Chuyuan Fu , Keerthana Gopalakrishnan , Karol Hausman , Alex Herzog , Daniel Ho , Jasmine Hsu , Julian Ibarz , Brian Ichter , Alex Irpan, Eric Jang , Rosario Jauregui Ruano , Kyle Jeffrey , Sally Jesmonth , Nikhil J Joshi , Ryan Julian , Dmitry Kalashnikov , Yuheng Kuang , Kuang-Huei Lee , Sergey Levine , Yao Lu , Linda Luu , Carolina Parada , Peter Pastor , Jornell Quiambao , Kanishka Rao , Jarek Rettinghouse , Diego Reyes , Pierre Sermanet , Nicolas Sievers , Clayton Tan , Alexander Toshev , Vincent Vanhoucke , Fei Xia , Ted Xiao , Peng Xu , Sichun Xu , Mengyuan Yan , Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. https://sites.research.google/palm-saycan

5. Peter Chen. AI Robotics' "GPT Moment" is near. https://techcrunch.com/2023/11/10/ai-robotics-gpt-moment-is-near/

6. Brian Heater. Robotics Q&A with UC Berkeley's Ken Goldberg. https://techcrunch.com/2023/12/16/robotics-qa-with-with-uc-berkeleys-ken-goldberg/

7. Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. https://deepmind.google/discover/blog/rt-2-new-model-translates-vision-and-language-into-action/

8. Dhruv Shah, Michael Equi, Blazej Osinski, Fei Xia, Brian Ichter, Sergey Levine. Navigation with Large Language Models: Semantic Guesswork as a Heuristic for Planning. https://sites.google.com/view/lfg-nav/

9. Mario Viviani. AI Room Generator Extension Sample. https://github.com/NVIDIA-Omniverse/kit-extension-sample-airoomgenerator

10. Brian Heater. Robotics Q&A with Meta's Dhruv Batra. https://techcrunch.com/2023/12/02/robotics-qa-with-metas-dhruv-batra/

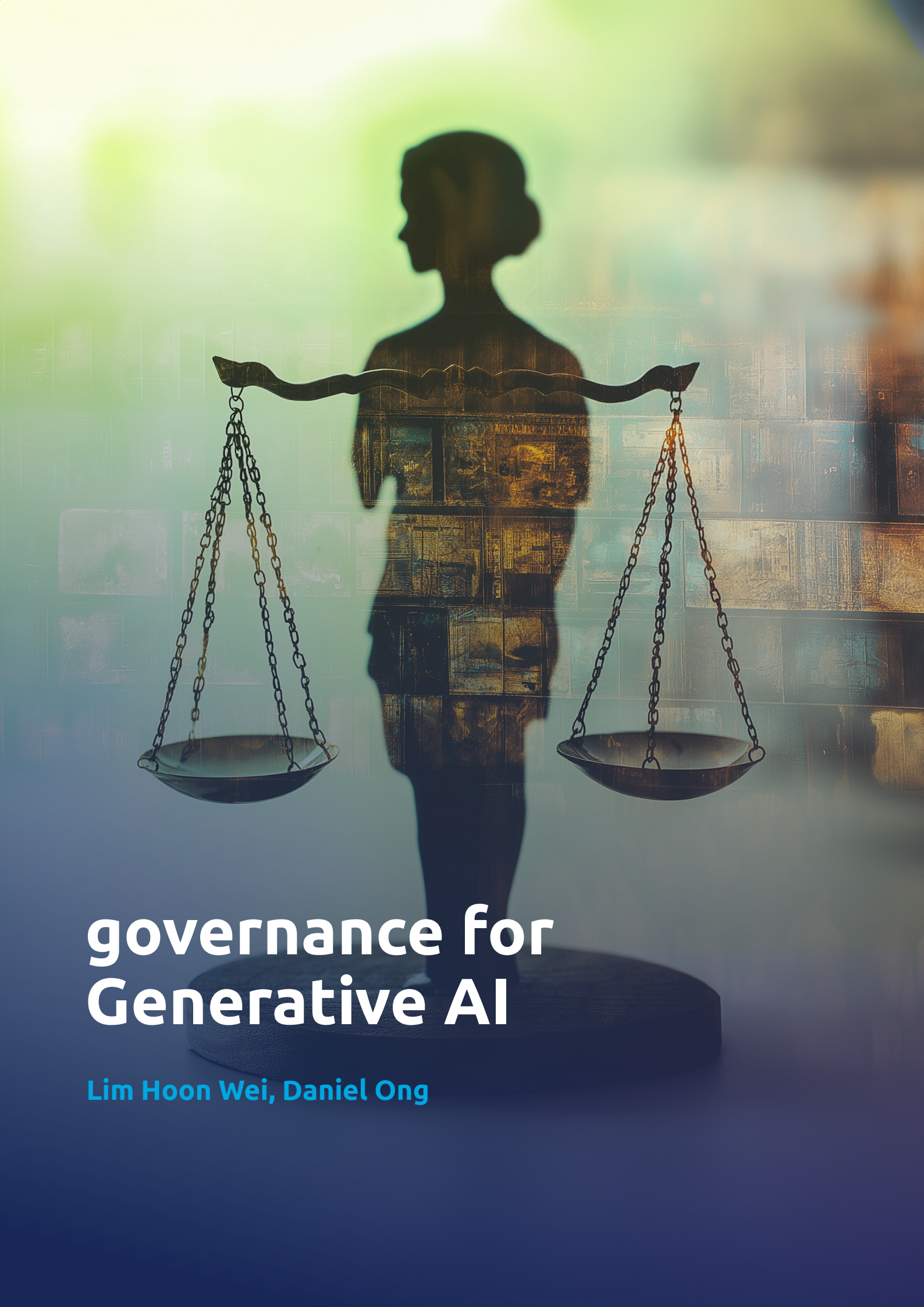
11. Brian Heater. Robotics Q&A with Nvidia's Deepu Talla. https://techcrunch.com/2023/12/16/robotics-qa-with-nvidias-deepu-talla/

12. Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas,Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian,Sean Kirmani, Isabel Leal, Edward Lee, Sergey Levine, Yao Lu, Sharath Maddineni, Kanishka Rao,Dorsa Sadigh, Pannag Sanketi, Pierre Sermanet, Quan Vuong, Stefan Welker, Fei Xia, Ted Xiao,Peng Xu, Steve Xu, Zhuo Xu.
AutoRT: Embodied Foundation Models for Large Scale Orchestration of Robotic Agents. https://auto-rt.github.io

13. Figure AI Inc. Figure Raises $675M at $2.6B Valuation and Signs Collaboration Agreement with OpenAI. https://www.prnewswire.com/news-releases/figure-raises-675m-at-2-6b-valuation-and-signs-collaboration-agreement-with-openai-302074897.html

14. 1x. 1X Raises $23.5M in Series A2 Funding led by OpenAI. https://www.1x.tech/discover/1x-rasies-23-5m-in-series-a2-funding-led-by-open-ai

15. Nvidia. Nvidia Announces Project GR00T Foundation Model for Humanoid Robots and Major Isaac Robotics Platform Update. https://nvidianews.nvidia.com/news/foundation-model-isaac-robotics-platform



governance for Generative AI

Lim Hoon Wei, Daniel Ong

governance for Generative AI

Lim Hoon Wei, Daniel Ong

Introduction

Generative AI (GenAI) has emerged as a groundbreaking technology with immense potential to revolutionise various aspects of our society. From enabling digital transformation in both large corporations and small organisations, to enhancing creativity and productivity, GenAI, such as ChatGPT, has opened new avenues for innovation and progress. However, while we celebrate the possibilities offered by this technology, it is crucial to acknowledge that its unrestricted use can give rise to significant risks and unintended consequences.

In the context of unrestricted use GenAI, we delve into the multifaceted nature of GenAI, particularly ChatGPT, and the need for an appropriate governance framework to ensure responsible and beneficial deployment. We explore the inherent risks associated with the unbridled use of ChatGPT and shed light on potential challenges that may arise if adequate precautions are not taken.

Recognising the importance of a balance between harnessing the power of GenAI and safeguarding against its potential pitfalls, we aim to equip individuals, organisations, and policymakers with a comprehensive understanding of the risks involved. Moreover, we provide a set of general guidelines that can serve as a starting point for mitigating these risks to ensure the responsible and ethical use of ChatGPT.

Through the establishment of effective governance measures, we can foster an environment where ChatGPT and similar technologies can thrive, to the benefit of society.

Benefits and risks of Generative AI

The rapid growth and use of GenAI in many applications is due to its potential benefits in multiple dimensions:

- Increased productivity: GenAI can help people to be more productive by automating tasks that are time-consuming or repetitive. For example, GenAI can be used to automate email responses for simple enquiries, schedule appointments, and summarise interactions.
- Enhanced creativity: GenAI can help people to be more creative by providing them with innovative ideas and inspiration. For example, it can be used to generate new product designs, musical compositions, and works of art.
- There is a trade-off between productivity and safety. Implementing safeguards to prevent the unethical use of ChatGPT can reduce the tool's productivity. For example, if ChatGPT is trained to identify and avoid generating harmful content, it may also generate less creative or less interesting content.
- Personalised experiences: GenAI can be used to create personalised experiences for users. For example, it can be used to recommend products to customers, generate tailored news feeds, and create custom educational content.

However, the use of GenAI can be a double-edged sword as the technology also poses several potential risks [IMDA23a, Gartner23]:

- Model explainability and interpretability: GenAI models can be very complex and difficult to understand. This can make it difficult to determine how they are making decisions and to identify potential biases or errors.
- Data quality and bias: GenAI models are trained on large datasets of existing content. If these datasets are biased, the models will also be biased. This can lead to the generation of content that is biased and discriminatory.
- Safety and security: GenAI models can be used to create malicious content, such as deepfakes or spam. It is important to develop safeguards to prevent the misuse of Generative AI models.
- Copyright infringement: GenAI models can be used to create works that are substantially similar to existing copyrighted works. If the model has been trained on copyrighted data, without permission, then the use of the model to create new works could be considered copyright infringement.
- Legal and ethical implications: The use of GenAI raises a number of legal and ethical issues, such as copyright, privacy, and liability. It is important to develop clear guidelines for the responsible use of GenAI.
- Economic and social impact: The widespread use of GenAI could have a significant impact on the economy and society. For example, it could be used to automate tasks that are currently performed by humans, which could lead to job losses. Additionally, GenAI could be used to create new products and services that could have a disruptive impact on existing industries.

In addition to these challenges, it is also important to be aware of the potential for GenAI to be used for malicious purposes. For example, it could be used to create fake news articles, social media posts, or even videos. This content could be used to spread misinformation and disinformation, which could have a negative impact on society.

Arising challenges

How do we strike a balance between enhanced productivity and the unethical use of ChatGPT?
There are a few reasons why such a balancing act is challenging:

- ChatGPT is a powerful tool that can be used for both good and bad. On the one hand, ChatGPT can be used to automate tasks, generate creative content, and improve communication. On the other hand, it can also be used to generate harmful or misleading content, such as fake news or propaganda.

- It can be difficult to distinguish between ethical and unethical use. In some cases, it may be clear that ChatGPT is being used unethically. For example, if someone is using ChatGPT to generate fake news, it is clear that they are using the tool for malicious purposes. However, in other cases, it may be more difficult to determine whether ChatGPT is being used unethically. For example, if someone is using ChatGPT to generate creative content, it may be difficult to determine whether the content is harmful or misleading.

- There is a trade-off between productivity and safety. Implementing safeguards to prevent the unethical use of ChatGPT can reduce the tool's productivity. For example, if ChatGPT is trained to identify and avoid generating harmful content, it may also generate less creative or less interesting content.

- It is difficult to regulate the use of innovative technologies. As new technologies emerge, it can be difficult to develop effective regulations to govern their use. This is especially true for technologies like ChatGPT, which are constantly evolving.

As a result of these challenges, it is non-trivial to strike a balance between reaping productivity gains and the unrestricted or unethical use of ChatGPT. More work needs to be done to mitigate the risks and ensure that the tool is used in a responsible manner.

Does OpenAI have the legal right to use the personal information of individuals to train ChatGPT?

According to Gartner, data privacy is the biggest concern based on a survey of 713 IT Executives. This is summarised in Figure 1 below.

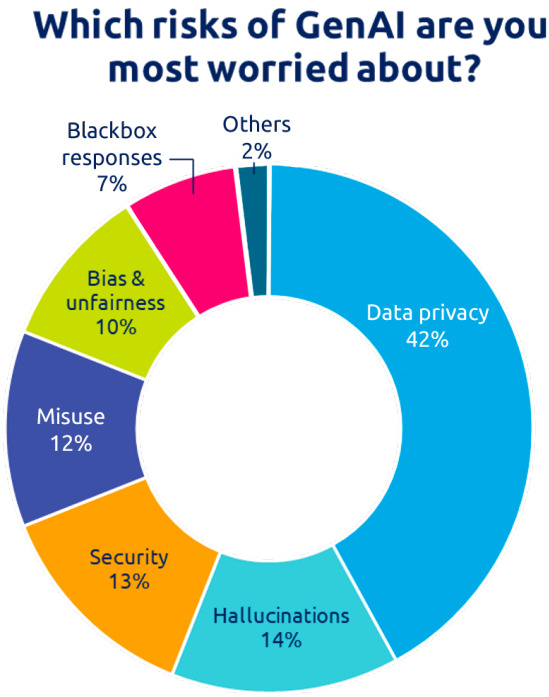


Figure 1. Risks associated with Generative AI according to 713 IT Executives (source: Gartner, Aug 2023)

In general, OpenAI is likely to have the legal right to use personal information to train ChatGPT if the data is collected and used in compliance with applicable laws and if users have consented to this use. If OpenAI collects personal information from users who have agreed to its terms of service, which state that the data may be used to train ChatGPT, then OpenAI is likely to have the legal right to use that data for this purpose.

However, there are some potential legal limitations on OpenAI's use of personal information to train ChatGPT. Some jurisdictions have laws that restrict the collection and use of certain types of personal information, such as sensitive personal data or data pertaining to children.

Additionally, some jurisdictions have laws that give individuals the right to opt out of having their personal data used for certain purposes, such as for training AI models.

For example, ChatGPT was banned by the Italian data protection authority at the start of April 2023 over privacy concerns. The Italian data protection authority, also known as Garante, temporarily restricted the chatbot and launched a probe over the artificial intelligence application's suspected breach. As Garante had accused OpenAI of failing to check the age of ChatGPT's users (who are supposed to be aged thirteen or above), OpenAI said it would offer a tool to verify users' ages in Italy upon sign-up. OpenAI explained that it would also provide a new form for European Union users to exercise their right to object to its use of their personal data to train its models. Access to the ChatGPT chatbot has since been restored in Italy [McCallum23].

Responsible AI through governance

Different jurisdictions may adopt different policies and approaches

The Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) published guidelines in July 2023, allowing limited use of Generative AI (GenAI) in elementary, junior high, and high schools. The guidelines are intended to help teachers and students understand the characteristics of the technology, while imposing some limits due to fear of copyright infringement, personal information leaks, and plagiarism [Kyodo23].

The guidelines state that Generative AI can be used for educational purposes, such as:

- Generating creative text formats of text content, like poems, code, scripts, musical pieces, email, letters, etc.
- Translating languages
- Writing various kinds of creative content
- Answering questions in an informative way
- Helping with research and problem-solving

However, the guidelines also caution that Generative AI should not be used for the following purposes:

- Plagiarism: Passing off AI-assisted schoolwork as one's own will be deemed cheating.
- Copyright infringement: Using Generative AI to create content that infringes on the copyright of others is prohibited.
- Writing various kinds of creative content.
- Personal information leaks: Using Generative AI to generate content that contains personal information about others is prohibited.
- Hate speech and misinformation: Using Generative AI to generate content that is hateful, misleading, or discriminatory is prohibited.

The guidelines also state that elementary school students younger than age 13 can use AI, to a certain extent, under the guidance of teachers.

The MEXT's decision to allow limited use of Generative AI in schools was a significant step, as Japan became one of the first countries in the world with such a government policy. The guidelines are still in the initial stages, and it remains to be seen how they will be implemented in practice. However, the move is a sign that Japan is recognising the potential of Generative AI to transform education.

On the other hand, the Australian Federal Government opened an inquiry last year into the use of Generative AI within its education system [Parliament23]. The Australian Federal Government's inquiry into the use of Generative AI in education is a significant development, as it was one of the first governments in the world to take such a step.

The inquiry will examine the following key areas:

- The potential benefits and risks of using Generative AI in education
- The ethical implications of using Generative AI in education
- The best ways to use Generative AI to improve education outcomes for all students

Schools in Australia have now officially rolled out the use of artificial intelligence bot ChatGPT. The decision to allow ChatGPT in schools was made after a consultation with teachers, students, parents, and education experts. The consultation found that there is a strong demand for the use of ChatGPT in schools, and that the potential benefits of using this technology outweigh the risks [Cassidy23, Beaumont24].

The Australian government is working with OpenAI to develop guidelines for the responsible and ethical use of ChatGPT in schools. The guidelines will cover areas such as:

- Data privacy and security
- Copyright and intellectual property
- Plagiarism and academic integrity
- Bias and discrimination

Where to start?

To address the ethical and societal concerns surrounding AI, including Generative AI, it is essential to develop and operationalise an effective governance framework within an organisation. These frameworks should be designed to ensure that AI is used in a responsible and ethical manner.

Some of the overarching principles of effective AI governance include:

- **Transparency:** AI systems should be designed in a way that is transparent and accountable. This means that it should be possible to understand how these systems work and to identify the biases that they may contain.
- **Fairness:** AI systems should be designed in a way that is fair and does not discriminate against any particular group of people. AI systems should be trained on datasets that are representative of the population, and they should be used in a way that does not perpetuate existing biases.
- **Accountability:** There should be clear lines of accountability for the development and use of AI systems, and it should be clear who is responsible for ensuring that these systems are used in a responsible and ethical manner.
- **Safety:** AI systems should be designed in a way that is safe and does not pose a risk to human safety or security. AI systems should be tested for potential vulnerabilities, and they should be equipped with safeguards to prevent misuse.
- **Privacy:** AI systems should respect user privacy. They should not collect or use any personal data without the user's consent.

Once the overarching AI principles have been defined, a governance framework that sets out how these principles will be put into practice needs to be developed. This framework should include policies and procedures for the development, deployment, monitoring, and evaluation of AI systems.

Here are some examples of key elements of an AI governance framework:

- AI ethics committee: An AI ethics committee should be established to review all AI projects and ensure that they are aligned with the organisation's AI principles.
- AI risk assessment: All AI projects should be subjected to a risk assessment to identify and mitigate any potential risks.
- AI monitoring: AI systems should be monitored to ensure that they are performing as expected and that they are not having any unintended negative consequences.
- AI impact assessment: AI systems should be regularly assessed to measure their impact on individuals, society, and the environment.

The US Department of Commerce National Institute of Standards and Technology NIST AI Risk Management Framework [NIST23] and Singapore's Infocomm Media Development Authority's Model AI Governance Framework [IMDA23b] provide useful guidelines and further details on how an effective governance framework might look.

AI risk and impact assessments should be conducted regularly to identify and mitigate any potential risks. These assessments should consider the following factors:

- The type of AI system: Some AI systems are riskier than others. For example, AI systems that are used to make critical decisions, such as for medical diagnoses or legal decision-making, pose higher risks than AI systems that are used for entertainment or marketing purposes.
- The data used to train the AI system: AI systems are trained on data and the quality and bias of the data can have a significant impact on the performance of the system. It is important to assess the risk of bias in the data used to train AI systems.
- The potential impact of the AI system: AI systems can have a positive or negative impact on individuals, society, and the environment. It is important to assess the potential impact of AI systems before they are deployed.

By conducting regular AI risk and impact assessments, organisations can ensure that AI systems are used in a safe and responsible manner.

Summary

GenAI can be used to create a wide variety of content, including text, code, images, and videos. This content can be used for good or malicious purposes, so it is important to have safeguards in place to prevent GenAI from being used maliciously.

Some recommended practices that organisations should consider when adopting a responsible approach to AI, in particular with the rise of GenAI and ChatGPT, are as follows:

- **Approve only a certain set of use cases:** Decide which use cases for ChatGPT are acceptable and which are not. For example, an organisation might allow ChatGPT to be used to generate personalised learning materials for students, but not to create fake news articles or deepfakes.

- **Avoid usage of certain input data:** ChatGPT is trained on a massive dataset of text and code which could contain personal information, confidential data, or biased content. It is important to avoid using ChatGPT to process any type of data that is sensitive or that could be used to harm or discriminate against individuals or groups.

- **Verify the accuracy of output from ChatGPT:** ChatGPT is a powerful language model, but it is not perfect. It is important to verify the accuracy of any output from ChatGPT before using it. This is especially important for high-stakes applications, such as medical diagnosis or legal decision-making.

- **Consider legal issues:** When using ChatGPT, it is important to be aware of the relevant laws and regulations. It is critical to ensure that no privacy laws are violated during the data collection process and that no copyrights are infringed through the generation of content.

In addition to these recommendations, it is also important to be transparent about how ChatGPT is used. Organisations should inform users about what data is being collected, how it is being used, and what they can expect from the output of the model. It is also important to have a process in place for users to report any problems or concerns they have about the use of ChatGPT.

By taking these steps, we will go a long way in ensuring that ChatGPT is used in a responsible and ethical way.

References

[IMDA23a] Generative AI: Implications for Trust and Governance.
The Infocomm Media Development Authority (IMDA) and Aicadium, 7 Jun 2023.

[Gartner23] Gartner Experts Answer the Top Generative AI Questions for Your Enterprise.
https://www.gartner.com/en/topics/generative-ai. Accessed on 24 Oct 2023.

[McCallum23] S. McCallum. ChatGPT accessible again in Italy. BBC News,
https://www.bbc.com/news/technology-65431914. 28 Apr 2023.

[Kyodo23] Japan publishes guidelines allowing limited use of AI in schools. Kyodo News,
https://english.kyodonews.net/news/2023/07/ac1ce46ce503-japan-publishes-guidelines-allowing-limited-use-of-ai-in-schools.html. 4 Jul 2023.

[Parliament23] Inquiry into the use of generative artificial intelligence in the Australian education system.
Parliament of Australia, https://www.aph.gov.au/Parliamentary_Business/Committees/House/Employment_Education_and_Training/AIineducation. Accessed on 25 Oct 2023.

[Cassidy23] C. Cassidy. Artificial intelligence such as ChatGPT to be allowed in Australian schools from 2024.
The Guardian, https://www.theguardian.com/australia-news/2023/oct/06/chatgpt-ai-allowed-australian-schools-2024. 6 Oct 2023.

[Beaumont24] D. Beaumont, A New Era of Technology: ChatGPT within Australian Schools.
The City Journal, https://thecityjournal.net/news/a-new-era-of-technology-chatgpt-within-australian-schools/. 24 May 2024.

[NIST24] NIST AI Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1,
https://doi.org/10.6028/NIST.AI.600-1. Jul 2024.

[IMDA24] IMDA Model AI Governance Framework, https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf. 30 May 2024.

Generative AI solutions for contact centre operations

Vijay Nayudu

Generative AI solutions for contact centre operations

Improving productivity and compliance management of contact centre agents with a new generation of AI Assistants

Vijay Nayudu

Customer contact centres play a pivotal role across a wide array of industries, including telecommunications, banking, and the public sector. These centres serve as the front line for customer interactions, addressing queries, resolving issues, and enhancing customer satisfaction. The contact centre industry is not just widespread but also deeply globalised. Many organisations have outsourced and offshored their customer service functions to leverage cost efficiencies and access to skilled labour in various regions.

As of today, the global contact centre industry employs over 17 million people worldwide and represents a market worth US$30 billion in 2023. This expansive industry continues to evolve rapidly, driven by technological advancements, with AI poised to be a transformative force.

In this article, we explore how AI impacts contact centres in two important ways: 1) improving the productivity of the contact centre agent, by allowing the agent to not only do existing tasks better in less time, but also take on new value-adding tasks; and 2) enhancing the effectiveness of organisation to ensure the agents comply with organisation-specific policies and industry rules and regulations.

Boosting productivity

Contact centre staff perform a variety of essential tasks that ensure smooth and efficient customer interactions. Their responsibilities include answering inbound calls, handling customer complaints, processing orders and payments, and providing detailed information about products and services. In addition to managing customer accounts and scheduling appointments, they often respond to emails and live chat messages, ensuring that all customer queries are addressed promptly. They also follow up on customer requests and conduct satisfaction surveys to gather feedback for continuous improvement. Adhering to scripts and compliance guidelines, particularly in regulated industries, ensures consistency and adherence to legal standards.

With AI's growing presence, many of these routine tasks are expected to become automated, allowing staff to focus more on complex, value-added activities. The impact of AI would vary, depending on the nature of the call centre operations to begin with.

In a next-gen contact centre scenario that we envisioned with a group of Telcos, GenAI acts as an assistant to the call centre agent across the lifecycle of a call:

- Pre-call: Assist the agent with personalised questions to validate the identity of the caller
- During the call: To provide technical resolutions from historical cases and surface relevant promotions to the agent
- Post-call: Automate capture of summaries and downstream actions

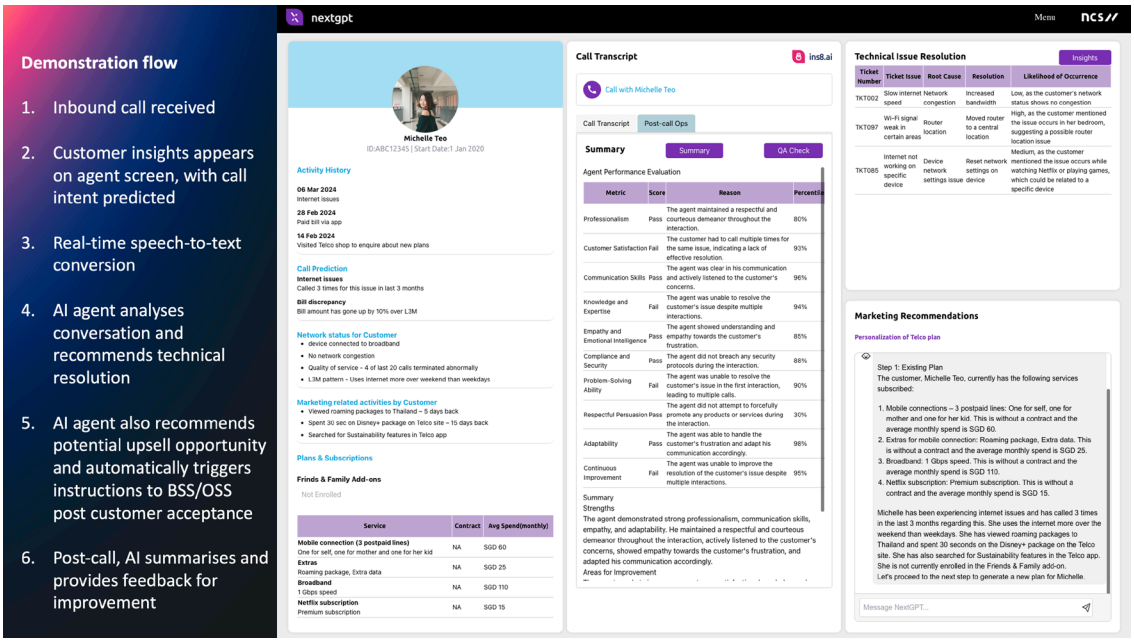


Figure 1. GenAI as an assistant to the call centre agent

In a recent implementation with a 250-strong call centre in Singapore, GenAI was applied successfully to 2 key areas:

- First, the training of new call centre agents. Compared to manual trainers, the AI training assistant always has access to up-to-date knowledge base as training content, is able to produce dynamic simulation scenarios to enhance the learning experience, and can be customised to take into account the knowledge level of the trainee. As a result of the AI training assistant, the onboarding time was reduced by 14%.

- Second area is automating the after-call work. The AI summariser generates post-call summaries in standardised format for follow-ups. This has helped to cut more than half of the after-call work time, freeing up agents to take on more calls. The agent can now handle more calls, up from average of 32 to 40 per day.

The underlying technology products and services used to modernise this contact centre included AWS Bedrock cloud services, Anthropic's Claude Instant LLM and NCS' proprietary speech-to-text engine called Ins8.ai. With the successful implementation, the technology architecture and deployment roadmap now serve as a reference template for other similar contact centres globally

Enhancing compliance

In the highly regulated world of financial services, banks, insurance companies and other advisory firms face significant challenges in ensuring compliance and preventing misconduct by their customer service agents or relationship managers (RMs). Manual and error-prone monitoring processes often lead to delayed detection of non-compliant behaviour, exposing banks to risks and potential penalties.

Banks and insurance companies face multiple hurdles in ensuring compliance within their organisations including a high volume of interactions, staff turnover, frequent product and policy updates, and the emergence of new regulations. The current manual approaches to compliance are insufficient to keep pace with the scale and speed of change.

Banks require large compliance teams to manually monitor just a small percentage of calls, resulting in significant latency and less-than-comprehensive oversight of non-compliant behaviour. Recent cases of personal bankers deceiving customers resulted in financial loss and damage to the banks' reputations, and highlight the limitations of the existing approach to compliance, such as the case reported by The Straits Times on 21 Aug 2024 wherein MAS issued a 9 year prohibition order against a large ASEAN bank's relationship manager for misconduct. Earlier this year, in Jan 2024, Bloomberg reported on the Credit Suisse AT1 bond fallout hitting a large Indian bank, where customers alleged that the bank's relationship manager mis-sold the bonds by saying they were secure and failing to explain the risks.

Leveraging advanced language models and automated hyperlocal speech recognition, NCS developed the ***Compliance AI Assistant*** to enable real-time monitoring, actionable insights, and enhanced performance management. This innovative solution addresses the problem statement "Is my Relationship Manager compliant with both internal and external regulatory policies and guidelines?' and utilises real-world conversations, bank policies, ethical guidelines, and regulations (such as those of the Financial Advisers Act/MAS, in Singapore) as inputs. By combining these inputs with the power of GenAI, the ***Compliance AI Assistant*** provides real-time alerts on non-compliance, generates comprehensive reports, and facilitates targeted training plans for RMs, enabling a holistic view of RM performance (Refer to Figures 2 and 3).

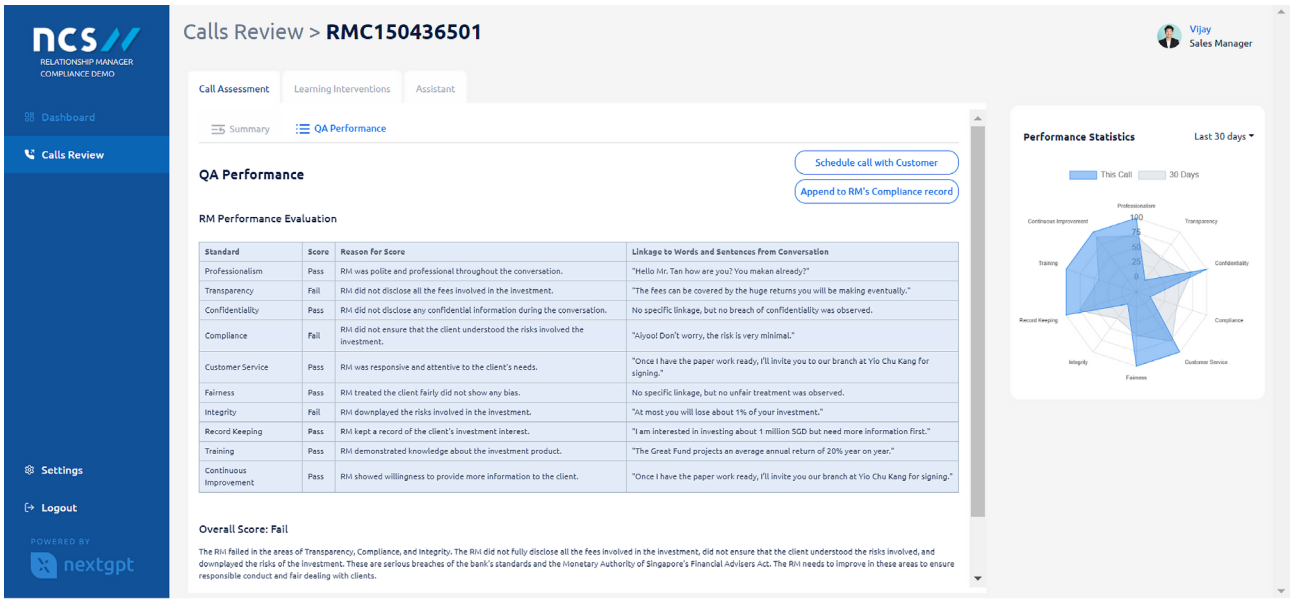


Figure 2. Automated compliance check in near real-time

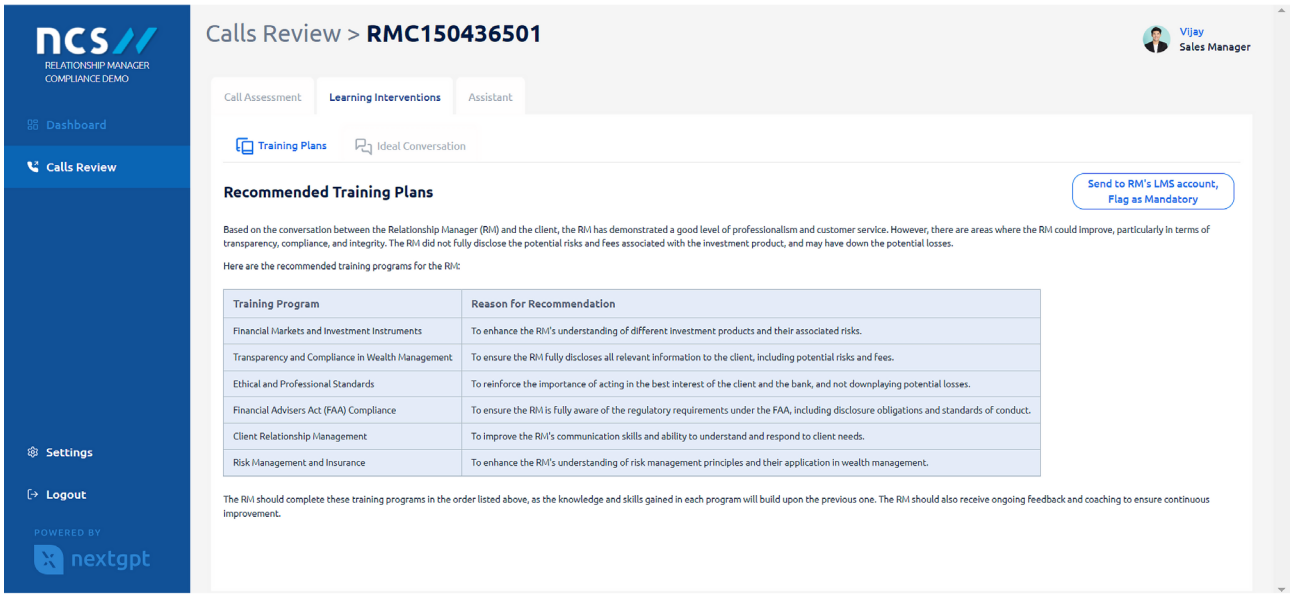


Figure 3. Recommended relevant learning interventions to improve agent performance

Implementing the ***Compliance AI Assistant*** and integrating it within a bank's business processes/IT environment can lead to significant improvements in compliance metrics. The percentage of calls monitored can be increased to 100% from the current 5-10% standard, instantly reducing the 3-6 months latency observed today. Furthermore, the solution can reduce the number of manhours by 80-90%, thus streamlining the compliance monitoring process and making it more efficient.

Conclusion

The integration of AI, particularly Generative AI, into contact centre operations offers transformative benefits by significantly enhancing both productivity and compliance management. On the productivity front, AI-driven solutions enable agents to perform routine tasks more efficiently, reducing time spent on repetitive activities and freeing them to focus on more complex, value-added interactions. By automating processes such as live speech-to-text transcription, post-call summaries, and personalised training, AI allows agents to handle more calls, improve customer satisfaction, and take on new responsibilities that contribute to business growth.

In terms of compliance management, AI is a game-changer for industries like banking and insurance, where regulatory adherence is critical. The ***Compliance AI Assistant***, for instance, provides real-time monitoring and insights, ensuring that every interaction is aligned with both internal policies and external regulations. This not only mitigates risk but also streamlines the compliance process by automating what was once a labour-intensive task. With AI, contact centres can monitor 100% of interactions, significantly reducing latency and minimising the likelihood of non-compliance. In essence, AI empowers organisations to enhance productivity while maintaining rigorous compliance standards, driving both efficiency and effectiveness in contact centre operations.