RBB| Economics

Generative AI: an assessment of competitive dynamics in the Asia-Pacific region[1]

RBB Economics, February 2026

[1] The preparation of this paper has been supported by the Computer & Communications Industry Association. The opinions expressed are those of RBB Economics and do not necessarily represent the views of CCIA or of its members.

1 Executive summary

This report assesses recent trends and competitive dynamics in Generative AI (**GenAI**), focusing on the Asia-Pacific (**APAC**) region. Our analysis examines possible competition concerns raised by regulators and evaluates whether structural features in APAC markets support these theories of harm or create conditions conducive to competition.[2] To support our discussion, we cover six APAC case studies which provide direct evidence on the nature of competitive dynamics across multiple levels of the GenAI value chain.

Our analysis points in one clear direction: while it remains prudent for regulators to monitor the sector carefully, concerns about concentration and adverse consumer outcomes have not materialised to date. Put simply, as of now, GenAI markets in the APAC region show characteristics consistent with healthy competition and appear to be displaying considerable dynamism with continued high levels of investment, entry and innovation.

1.1 The APAC GenAI landscape

GenAI refers to artificial intelligence systems that can create new content (text, images, code, audio and video) in response to user prompts. Unlike traditional software that follows fixed rules, GenAI generates outputs based on patterns learned from vast amounts of training data.

Our report focuses on the following layers of the supply chain:

- **Foundation model layer.** Foundation models (**FMs**) serve as the core technology underpinning GenAI. These models analyse patterns in large training datasets and learn to predict the next output in a sequence.[3] Over multiple training stages, this allows the model to generate content in response to more complex user queries. Given the large data requirements, FMs require substantial computing resources and technical expertise to train. FMs are increasingly multimodal, able to interpret and generate text, image, audio, video and code. FMs used across APAC include text models such as OpenAI's GPT, Anthropic's Claude, Google's Gemini, Alibaba's Qwen, Baidu's ERNIE, Tencent's Hunyuan, and Naver's HyperClova X; image models such as Open AI's DALL-E, Google's Nano Banana, and DeepSeek's Janus; and code-specialised models such as Mistral's Codestral and Alibaba's Qwen Coder.

- **Deployment platforms layer.** These platforms help downstream customers build and deploy GenAI applications. Common features include "model gardens" (catalogues that provide access to FMs from multiple providers through a single interface) and development tools which enable organisations to integrate, compare, and switch between these models. Some examples of deployment platforms include Google Vertex AI, Amazon Bedrock, Microsoft Azure AI, Baidu's Qianfan platform and Alibaba Cloud Model Studio.

- **Applications layer**. The application layer comprises products and services built on FMs to deliver GenAI capabilities to enterprises or customers. These may be standalone applications or GenAI features integrated into existing products. Developers integrate GenAI capabilities into their applications through API access (where the FM provider maintains the model and the developer pays for usage) or self-hosted deployment (where the developer downloads an open-source

[2] We focus this analysis on the FM, deployment platform, and application levels of the GenAI value chain. We do not comment extensively in this report on the input markets, such as cloud computing or specialised chips, only referring to them when relevant for the assessment of the state of competition in the downstream production of GenAI FMs, deployment platforms and applications.

[3] For example, a language model may learn to predict the next word in the sentence "the cat sat on the…". An image model may receive a partial picture of a table and learn to predict that the remaining part of the picture should show some chairs.

model or, less commonly, licenses a proprietary one, and tailors it to its needs). Fine-tuning models for specific tasks, such as document review, or code generation, can be a source of differentiation and allow firms to take advantage of their proprietary domain-specific data. Common examples include general purpose chatbots (such as, ChatGPT, Claude or ERNIE Bot), productivity assistants (such as, Notion AI and Microsoft Copilot), coding tools (such as Github Copilot and Claude Code), image generation (such as Adobe Firefly) and marketing copywriting tools (such as Jasper).

Taking a global view, the evidence suggests that competition at each layer remains effective with a broad range of competitors, diverse business models, and high levels of innovation. At the FM layer, for example, the evidence suggests that the number of competitors is growing and that (while some firms are clearly larger than others) no single operator has an unduly strong market position.[4]

From an APAC-specific perspective, our research suggests that there are a number of factors which could plausibly affect competitive dynamics. We note at the outset that the APAC region leads in terms of AI adoption.[5] The market opportunity for firms active in APAC is therefore large, encouraging firms to invest and innovate. The region also hosts a number of strong APAC-specific players across the GenAI value chain. Their presence indicates that local firms can compete effectively with globally established players (typically originating from the US) and compete with them globally, as growing numbers of Western developers rely on Chinese open-source FMs.[6]

Market features affecting competitive dynamics in the APAC region include:

- **Mobile-first distribution:** Consumers show a relative preference for mobile in the APAC region, incentivising suppliers to prioritise deployments that integrate well on mobile devices and favouring multimodal capabilities, given the prevalence of voice-based and image-based search on mobile devices.[7, 8]

- **Super app ecosystems:** Super apps such as WeChat, Alipay, LINE and Grab that bundle multiple services within a single platform are relatively more popular in the APAC region.[9] They therefore serve as a relatively more important route to market for GenAI suppliers.

- **Government investment and partnerships:** Government investment plays a larger role than in Western markets. China's projected 2025 AI capital expenditure (USD 84-98 billion) is predominantly from government funding (approximately 57%), with Japan, South Korea, and India also pursuing substantial public investment in domestic AI development.[10] In contrast, analysts estimate public sector investment accounts for less than 5% of total AI investment spending in Europe and the US.[11] These public sector investments often target sovereign infrastructure and domestic GenAI capabilities. In addition, various government initiatives attempt to lower barriers to GenAI adoption for key use cases such as creating local language datasets to aid the development of FMs with strong performance in

---

4 IOT Analytics, 'The leading generative AI companies', 4 March 2025. We do not comment on the validity of these estimates but merely note that these sources suggest there are many material competitors at the FM layer.

5 BCG, 'Asia Pacific Leads the World in AI Adoption While Grappling with Job Fears, BCG Survey Finds', 30 October 2025.

6 NBC News, 'More of Silicon Valley is building on free Chinese AI', 30 November 2025.

7 Globally, 42% of consumers had issued voice commands to their smartphone in the past six months as of 2022, compared to 14% issuing voice commands to laptops or desktops. This difference is even starker in APAC as more than 60% of consumers in Indonesia, China and India issued voice commands to their smartphone, compared to less than 30% using voice commands for laptops or desktops. YouGov, 'Global: On which devices are consumers using voice commands?', 6 September 2022.

8 Mobile devices were also the primary mode for image-based searches as of 2018, with 53% of image -based searches originating on mobile compared to 37% on laptop or desktop. DSI Spotlight, 'Visual Search Spotlight', October 2018.

9 Tencent, '2025 First Quarter Results', 14 May 2025.
CoinLaw, 'Alipay Statistics 2025: User Adoption, Transaction Volumes, and Technological Innovations', 16 June 2025.
Digital Marketing for Asia, 'Why is LINE the most popular social media app in Japan?', accessed November 2025.
Grab, 'Grab Reports Second Quarter 2025 results', 31 July 2025.
Grab, 'Where We Are', accessed November 2025.

10 TechWire Asia, 'China to deploy $98bn in AI investment this year amid US tech rivalry', 26 June 2025.

11 AWS, 'Artificial Intelligence Index Report 2025', 18 April 2025, Figures 4.3.10, 6.3.6, 6.3.12. Note that no AI-related grant data is available for Europe in the report.

native language tasks. It is well documented that Chinese competitors have had a significant impact on competition in the global FM market due to their effectiveness and cost advantages.[12]

- **Regulatory environment:** The APAC region encompasses a number of legal jurisdictions with differing approaches to data protection and AI oversight, and uneven but growing cross-border alignment.[13] This creates a fragmented regulatory environment that imposes asymmetric costs and compliance requirements for firms. This may contribute to variability in supplier presence across the region and requires commercial partnerships for market entry in some circumstances. Regulatory obligations are most pronounced in China where FM developers are required to obtain regulatory approval and data handlers are subject to strict regulations regarding the use and transfer of sensitive information. GenAI regulation in the remaining APAC countries is less stringent, but there is still varying data protection legislation, ranging from firm data sovereignty requirements in some Southeast Asian countries to lighter touch regimes such as in Australia and New Zealand. These requirements often lead organisations to maintain relationships with multiple cloud providers to ensure regional compliance, which in turn exposes them to multiple and diverse GenAI offerings.

1.2 Competition concerns examined

Competition authorities across APAC and internationally have identified potential risks at multiple layers of the GenAI value chain: for instance, FM markets might consolidate around one or a few dominant providers, vertically integrated firms might foreclose rivals' access to essential inputs, and platform-mediated advantages might create barriers to entry.[14] This report examines whether market conditions in APAC suggest that these risks are likely to materialise.

For analytical purposes, we group the concerns raised by competition authorities into five categories:

- **Vertical integration concerns:** When one company controls multiple layers of the value chain (for example, cloud infrastructure, FMs and applications), could they restrict competitors' access to essential inputs or favour their own services over rivals?

- **Platform-mediated advantages:** Could platform-level network effects (if platforms become more valuable as more people use them) or bundling practices (combining AI with other services) create self-reinforcing mechanisms, leading to market consolidation towards certain platforms, or high switching costs that lock-in users?

- **Concentration:** Could markets “tip”, with a small number of large players then benefitting from strong and entrenched market power that can harm consumers?

- **Agreements between firms:** Could partnerships with large developers transfer effective control without triggering merger review, or create exclusivity that forecloses rival providers?

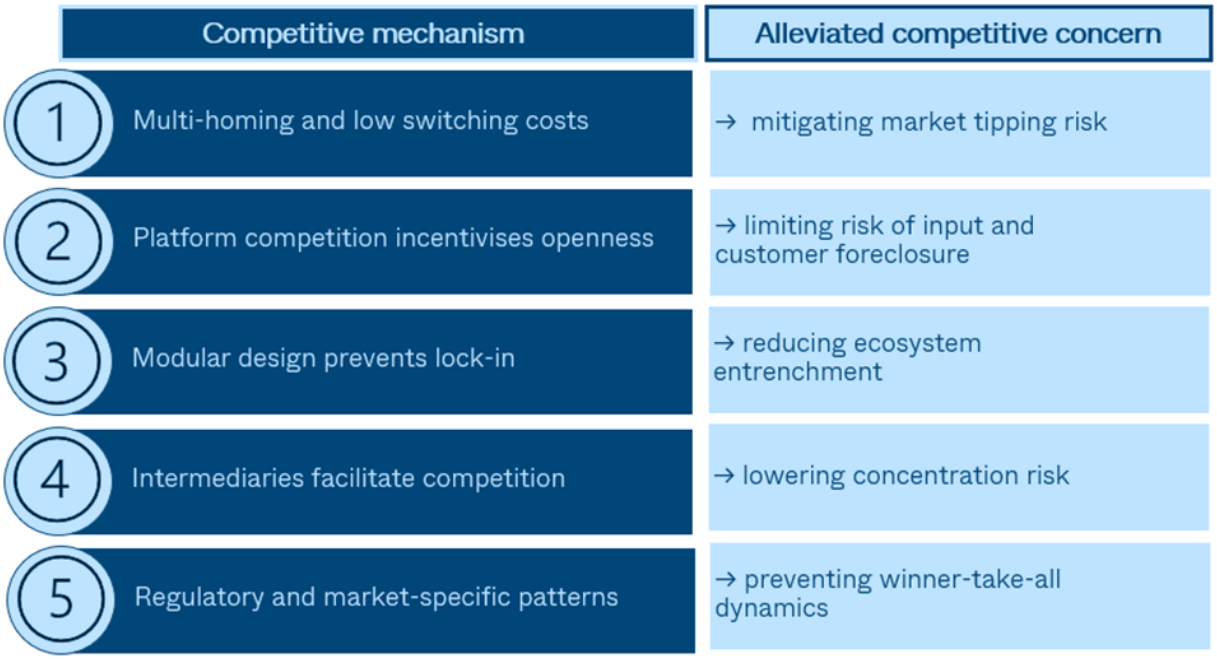
- **Mergers:** Could the acquisition of key talent or combination with an emerging rival, sidestep traditional merger review or remove potential competitors before they grow?

---

12 Chatham House, 'Low-cost Chinese AI models forge ahead, even in the US, raising the risks of a US AI bubble', 21 November 2025.
13 Microsoft, 'Asia Pacific's AI Leap: From Strategic Drive to Agentic Innovation', 6 November 2025.
14 KFTC, 'Generative AI and Competition', 18 December 2024.
JFTC, 'Report regarding Generative AI ver. 1.0', 6 June 2025.
TFTC, 'Competition Law Issues Related to Generative Artificial Intelligence, consultation paper', 7 September 2025
ACCC, 'Digital platform services inquiry, final report', 13 March 2025.
CCI, 'Market study on artificial intelligence and competition', September 2025.
NZCC, 'Navigating the rise of AI: Perspectives from a competition and consumer regulator', 7 July 2025.
OECD, 'Artificial Intelligence, Data and Competition – Note by Singapore', 29 May 2024.
RBB, 'Competitive Dynamics of Generative AI', 12 June 2025, section 3.

Our analysis examines whether structural features of APAC markets create conditions conducive to these concerns or, rather, support competition. Case studies spanning China's closed ecosystem, open markets (Japan, South Korea, India, Southeast Asia and Australia), and hybrid regulatory environments reveal five competitive dynamics that impact this assessment.

Figure 1: Overview of five competitive mechanisms at the FM, deployment platform and application layers



Source: RBB.

1.2.1 Multi-homing and low switching costs

Competition authorities have identified market tipping as a potential risk, where FM markets might consolidate around one or a few dominant providers. Economic theory suggests tipping is more likely when high switching costs and lock-in effects prevent organisations from changing providers, or when data feedback loops create self-reinforcing advantages where early leaders accumulate user data that enables them to develop better models and to improve them more effectively. We find no evidence currently of such market characteristics.

Multi-homing is prevalent across FM deployments. Enterprises routinely deploy multiple FMs simultaneously, either relying on different models for different tasks and use cases or multiple models for the same uses. Approximately 20% of APAC firms describe their AI development strategy as using multiple FMs.[15] This pattern extends across diverse organisational structures as illustrated in the case studies detailed in section 5.2. Super apps like Grab (Southeast Asia's ride-hailing platform) use different models for driver features versus merchant tools. Service integrators like Tata Consultancy Services (**TCS**) maintain partnerships across AWS, Google Cloud, Microsoft Azure and IBM watsonx to provide services to their clients. Large conglomerates like Samsung partner with Google Gemini globally, Baidu ERNIE in China and Microsoft Azure AI for customer service while developing internal Gauss models. Financial institutions like CBA (Australia's largest bank) operate simultaneous co-development relationships with AWS, Microsoft, Anthropic, and OpenAI. While multi-homing may occur because different models work better for different tasks, established relationships with multiple suppliers and technical know-how are likely to reduce barriers to switching compared to using a single supplier. The

[15] BCG, 'In the Race to Adopt AI, Asia-Pacific Is the Region to Watch', 11 March 2025, p 7. This may underestimate the true scale of multi-homing as a further 59% of respondents replied that they mostly worked with large US tech firms, which may include some firms that work with multiple US FMs or utilise non-US models for a small percentage of their GenAI workloads.

pace of innovation also suggests this multi-homing may continue as firms continue to test and retest which models perform best for different tasks.

The ability to switch is demonstrated by event studies. In July 2024, OpenAI blocked Chinese developers from using its API, giving them only two weeks to find alternatives before their applications would stop working. Chinese developers switched to domestic FMs (such as Baidu, Alibaba or Zhipu) or found ways to continue using OpenAI models through Microsoft's China-based partner 21Vianet. Competing providers immediately offered help packages to attract these developers, including millions of free usage tokens (the units measuring API usage) and technical support to assist with migration. Despite the tight deadline, industry reports indicated that applications continued operating rather than shutting down, with domestic providers gaining customers rather than the market contracting.[16] This episode illustrates three points. First, switching was feasible at relatively limited costs: API formats were similar enough to OpenAI's that developers could migrate without completely rewriting their applications, and providers actively reduced friction through support packages. Second, the fact that many developers chose domestic models rather than routing through 21Vianet to retain OpenAI indicates that Chinese providers had developed models with comparable performance, making them a competitive alternative rather than a fallback. Third, active competition amongst suppliers through low prices, switching incentives and high quality suggests that lock-in to any one supplier is unlikely.

Open-source models provide competitive alternatives that constrain proprietary providers' market power. For instance, DeepSeek published its innovative models with the full code, Alibaba's Qwen ranks first globally for downloads on the open-source platform Hugging Face, and Baidu released its latest ERNIE version for anyone to download and use freely.[17] Viable alternatives to paid proprietary models are experiencing widespread adoption in the region and beyond: 76% of Indian startups build primarily on open-source models and 36% of ASEAN startups use Meta's open-source Llama.[18] This adoption pattern shows open-source models offer sufficient capabilities for commercial applications, not just experimental use. When free alternatives offer comparable performance, proprietary providers cannot raise prices excessively or significantly decrease innovation efforts without losing customers to open-source options, maintaining competitive pressure even when proprietary providers hold strong market positions.

1.2.2 Platform competition limits incentives to engage in foreclosure

Competition authorities have raised concerns that vertically integrated firms operating across multiple value chain layers may restrict downstream rivals' access to upstream inputs (input foreclosure) or impede upstream rivals' access to downstream customers (customer foreclosure) (see section 3.1.1). We see no evidence that foreclosure concerns have materialised to date.

Observed behaviour across APAC suggests competitive incentives favour openness. Chinese technology platforms (such as Baidu, Alibaba, Tencent and ByteDance) operate cloud infrastructure, develop FMs, and control major consumer platforms, yet chose to integrate an independent competitor's FM into their consumer services alongside their own proprietary models. Tencent (which operates the WeChat messaging app serving 1.4 billion monthly users) is currently testing the incorporation of the independent DeepSeek model into its search features despite having developed its own FM, while Baidu (China's leading search engine) integrated DeepSeek alongside its proprietary model for advanced search queries.[19] Global cloud providers demonstrate similar openness: Amazon's model deployment platform,

---

16 STCN, '"Relocation" or "Going Global": OpenAI's Countdown to the Discontinuation of its Chinese APIs', 2 July 2024.

17 GitHub, 'DeepSeek GitHub repository', accessed November 2025.
Hugging Face, 'Qwen model statistics', accessed November 2025.
GitHub, 'ERNIE GitHub repository', accessed November 2025.

18 GenAI Fund, 'ASEAN GenAI Startup Report 2024', 15 September 2024, p 23.
Competition Commission of India, 'Market Study on Artificial Intelligence and Competition', September 2025, p 23.

19 Reuters, 'Tencent's Weixin app, Baidu launch DeepSeek search testing', 16 February 2025.
TechNode, 'Baidu Search integrates DeepSeek and Large Model ERNIE for advanced search', 17 February 2025.

Bedrock, offers over 100 FMs from multiple providers despite AWS developing its own Titan models, while Google Vertex AI distributes competitor models alongside Gemini.[20] Samsung (a South Korean conglomerate spanning semiconductors through consumer devices) distributes a competing Korean FM through its enterprise platform despite developing its own models internally.[21]

In our view, the intense nature of competition at the platform level ensures that there is no incentive to favour in-house models over best in class alternatives. WeChat competes with Alipay, Grab competes with Gojek, LINE competes with KakaoTalk. When users can compare AI capabilities across competing platforms, each platform must offer competitive AI features to attract and retain users. This competitive pressure means platforms cannot afford to restrict users to inferior proprietary models when rivals offer better AI through multi-homing: platforms must integrate the best-performing models available, whether developed internally or externally.

Enterprise development platforms demonstrate similar patterns. Samsung SDS positions FabriX as offering choice across competing providers (such as OpenAI, Naver, and Samsung's own models). Service integrators like TCS maintain partnerships with AWS, Google Cloud, Microsoft Azure, and IBM watsonx rather than channelling customers toward single suppliers. The consistency of multi-homing across both platform operators and integrators suggests that customers value FM variety and the flexibility to easily switch between these solutions depending on their specific requirements. Under such conditions, platforms that restrict access to proprietary models risk losing customers to competitors offering broader model choice.

1.2.3 Modular design enables partnerships without lock-in

Competition authorities have raised concerns that partnerships could effectively lead to problematic concentrations or incentive alignment without triggering merger review. In addition, exclusive vertical agreements with large purchasers could create scenarios where rivals are foreclosed from key routes to markets (see section 3.1.4). Our analysis focuses on deployment case studies, including instances where an organisation jointly develops a model or an application with an FM provider or a deployment platform. In certain circumstances, co-development and technical integration could result in high switching costs and supplier lock-in, yet observed behaviour shows that organisations can design flexible systems that enable them to maintain partnerships with multiple firms.

Simultaneous partnerships can provide technical integration taking advantage of the strengths of multiple suppliers and fostering competition between them. CBA operates simultaneous co-development relationships with four major FM providers: AWS (a decade-long partnership including AI development infrastructure and full data platform migration), Microsoft (Seattle-based engineers working directly with CBA teams), Anthropic (applications in fraud prevention and customer service enhancement), and OpenAI (exploratory fraud detection announced five months after Anthropic partnership expansion).[22] These partnerships involve more than API use, including joint engineering and supplier-specific architecture.

Systems designed for flexibility across multiple suppliers prevents customer lock-in. CBA selects amongst different models. The bank built an internal platform that sits between AI suppliers and applications and has implemented eleven guardrails on its GenAI models' inputs and outputs to enforce

---

[20] AWS, 'Amazon Bedrock Marketplace', accessed November 2025.
AWS, 'Overview of Amazon Titan models', accessed November 2025.
Google Cloud, 'Overview of Generative AI on Vertex AI', accessed November 2025.
[21] Samsung SDS, 'Samsung SDS to Lead Innovation in Hyperautomation with Generative AI', 14 September 2023.
[22] CBA, 'CommBank accelerates AI integration with major data migration to cloud', 4 June 2025.
CBA, 'CBA and Microsoft deepens Gen AI partnership', 11 March 2024.
CBA, 'CommBank expands strategic partnership with generative AI company, Anthropic', 14 March 2025
CBA, 'CommBank and OpenAI embark on Australia-first strategic partnership to advance AI solutions', 13 August 2025.

its security and compliance policies uniformly across providers.[23] This centralised platform enhances flexibility by creating a buffer between suppliers and CommBank's internal applications, as well as providing standardised tools to integrate products from multiple suppliers for future GenAI development projects. Indeed, CBA's platform is integrated with Amazon Bedrock and Amazon Bedrock Knowledge Bases which provide a standardised API for various models from leading AI companies.[24] This suggests firms can maintain collaborative relationships and supplier variety through design choices that preserve flexibility.

Parallel partnership for overlapping use cases demonstrates active supplier evaluation. CBA partners with both Anthropic and OpenAI for overlapping use cases in fraud and scam prevention. Further, CBA utilises multiple models within the same GenAI service when circumstances require. For example, when co-developing its CommBiz GenAI business banking messaging service with AWS, CBA chose to power it with multiple models (Claude and Cohere) rather than a single FM provider.[25] Further, Samsung partners with Baidu in China to use ERNIE for AI features that are powered by Gemini in other jurisdictions. This suggests direct substitutability between different FMs within certain use cases and leads to minimal switching costs for providers who multi-home.

1.2.4 Intermediaries facilitate competition

Market concentration concerns often focus on tipping dynamics whereby platform-level network effects and economies of scale favour single dominant providers (see section 3.2.2). Service integrators and platform aggregators acting as intermediaries between FM providers and end customers can either accelerate concentration (by channelling demand to preferred suppliers) or reduce it (by maintaining neutral multi-provider access). Observed behaviour shows intermediaries facilitating access to multiple providers, supporting model evaluation and switching, and providing compliance tools which support multi-homing across APAC markets.

Multi-platform integrators maintain neutral access encouraging competition between FM providers. TCS (India's largest IT services firm, with over 607,000 consultants across 55 countries) built partnerships with AWS, Google Cloud, Microsoft Azure, IBM watsonx, and Nvidia, positioning itself as a multi-platform integrator rather than a single-supplier channel.[26] The company's WisdomNext platform brings together GenAI capabilities with "intelligent evaluator bots" comparing available models to recommend optimal choices for specific workloads, facilitating competition between different FM providers for these customer deployments.[27] Actual deployments show TCS's enterprise customers actively choosing deployment platforms based on their own preferences: Singapore's FairPrice Group (a major retailer) uses Google Vertex AI, Wyndham Hotels uses AWS, while manufacturing plant operators rely on the TCS-Microsoft Azure partnership.[28]

---

[23] CBA, 'CommBank and AWS expand collaboration to deliver global best cloud and AI capabilities, enabling idea to production in six weeks', 4 February 2025.
CBA, 'Customer safety, convenience and recognition boosted by early implementation of Gen AI', 28 November 2024.

[24] CBA, 'CommBank and AWS expand collaboration to deliver global best cloud and AI capabilities, enabling idea to production in six weeks', 4 February 2025.
AWS, 'Amazon Bedrock', accessed November 2025.

[25] CBA, 'CommBank and AWS expand collaboration to deliver global best cloud and AI capabilities, enabling idea to production in six weeks', 4 February 2025.

[26] TCS, 'TCS Named Global Top Employer for 2025', 18 February 2025.
TCS, 'TCS Partners with Google Cloud to Integrate Gemini Enterprise for its Workforce and Customers', 14 October 2025.
TCS, 'TCS to Build New AI-Led Solutions for Business Transformation in Collaboration with Microsoft', 20 June 2025.
TCS, 'TCS launches new Generative AI practice in collaboration with AWS', 27 November 2023.
IBM, 'IBM Launches watsonx Code Assistant, Delivers Generative AI-powered Code Generation Capabilities Built for Enterprise Application Modernization', 26 October 2023.
TCS, 'TCS Launches NVIDIA Business Unit to Accelerate AI Adoption for Customers Across Industries', 24 October 2024.

[27] TCS, 'TCS Launches WisdomNext, an industry-first GenAI Aggregation Platform', 7 June 2024.

[28] FairPrice Group, 'FairPrice Group opens Store of Tomorrow at Punggol Digital District', 28 August 2025.
Google Cloud, 'FairPrice Group Unveils 'store of Tomorrow' Program with Google Cloud to Reimagine Its Retail Experiences and Operations', 5 June 2025.
TCS, 'TCS launches new Generative AI practice in collaboration with AWS', 27 November 2023.
TCS, 'TCS Launches 5G-Enabled Cognitive Plant Operations Adviser to Help Transform Plant Operations', 14 March 2023.

Independent orchestration platforms further reduce barriers and facilitate the comparison of FM performance even for companies with limited technical abilities. OpenRouter provides access to over 500 models from more than 60 providers through a single API, enabling developers to switch between providers by changing a single line of code.[29] LangChain is a development framework that acts as a universal adapter: developers build their application once, and LangChain handles the technical differences between FM providers, making it straightforward to switch models. It has achieved over 130 million downloads globally, supporting over 70 model providers, with 33% of open-source model users employing LangChain.[30] These orchestration tools make using multiple FMs and actively switching between them easier, even for less sophisticated deployers.

1.2.5 Regulatory and market-specific patterns support a wide range of competitors

Region-specific market characteristics prevent market tipping towards a single supplier at the FM, platform deployment and application deployment levels. This supports the existence of a wide range of competitors in the APAC region.

Regulatory diversity prevents uniform market access and region-wide dominance. Multicloud infrastructure reduces the risk of lock-in. Data sovereignty requirements in Indonesia, Vietnam, Singapore and China lead organisations to maintain multicloud architectures across multiple providers, both local and international. Almost 90% of APAC enterprises operate across multiple cloud platforms. Each platform offers access to different foundation model sets (AWS Bedrock to Claude, Llama and Mistral; Google Vertex AI to Gemini and Claude; and Alibaba Cloud Model Studio to Qwen, DeepSeek and Kimi), enabling organisations to test and compare models across providers. Regulatory restrictions also prevent region-wide dominance by any single provider. For example, Samsung partners with Google Gemini globally but substitutes Baidu ERNIE in China where Google services are blocked, providing identical functional capabilities (such as text summaries, translation and search) through different FMs depending on the regulatory requirements.[31]

Local language requirements and mobile-first applications favour domestic developers. Local expertise, market knowledge and data advantages enable domestic firms to compete effectively in these use cases. For example, Rakuten's AI 7B model achieved top performance among open Japanese Large Language Models (**LLMs**) and embedded a novel Japanese-language customised tokeniser algorithm.[32] Meanwhile, Naver's HyperClova X outperformed models developed by global players which had been adapted for Korean tasks.[33] However, this advantage has limits. Many applications can separate user-facing language handling from underlying processing: Rakuten's voice assistant combines its own Japanese model with OpenAI's Realtime API for speech capabilities, limiting the local language advantage to one component of the application rather than the whole system. Moreover, global developers can invest in local language capabilities if the market opportunity justifies it, and leading multilingual models are narrowing the performance gap on Korean benchmarks.

Cost-efficient training techniques and small language models (SLMs) lower entry barriers at the FM and deployment levels. Restrictions on chip exports to China may have helped spur a wave of model innovation focusing on efficient and streamlined model designs that reduce training costs, as demonstrated by the release of DeepSeek-R1. Separately, model developers are creating high-performing SLMs optimised for local deployment, such as Samsung's Tiny Recursion Model, a 7-million-

---

29 OpenRouter, 'About OpenRouter', accessed December 2025.
OpenRouter, 'Quickstart', accessed November 2025.
30 LangChain, 'LangChain's Second Birthday', 24 October 2024.
McKinsey, 'Open source technology in the age of AI', 22 April 2025, p 5.
31 Techwire Asia, 'Baidu AI Cloud partners Samsung Electronics China on generative AI smartphone', 29 January 2024.
Bloomberg, 'Samsung to Showcase Baidu's Ernie AI in Latest Galaxy Phones', 29 January 2024.
32 Rakuten Today, 'Rakuten's Open LLM Tops Performance Charts in Japanese', 20 April 2024.
33 Naver Cloud, 'HyperClova X Technical Report', 13 April 2024, Figure 1.

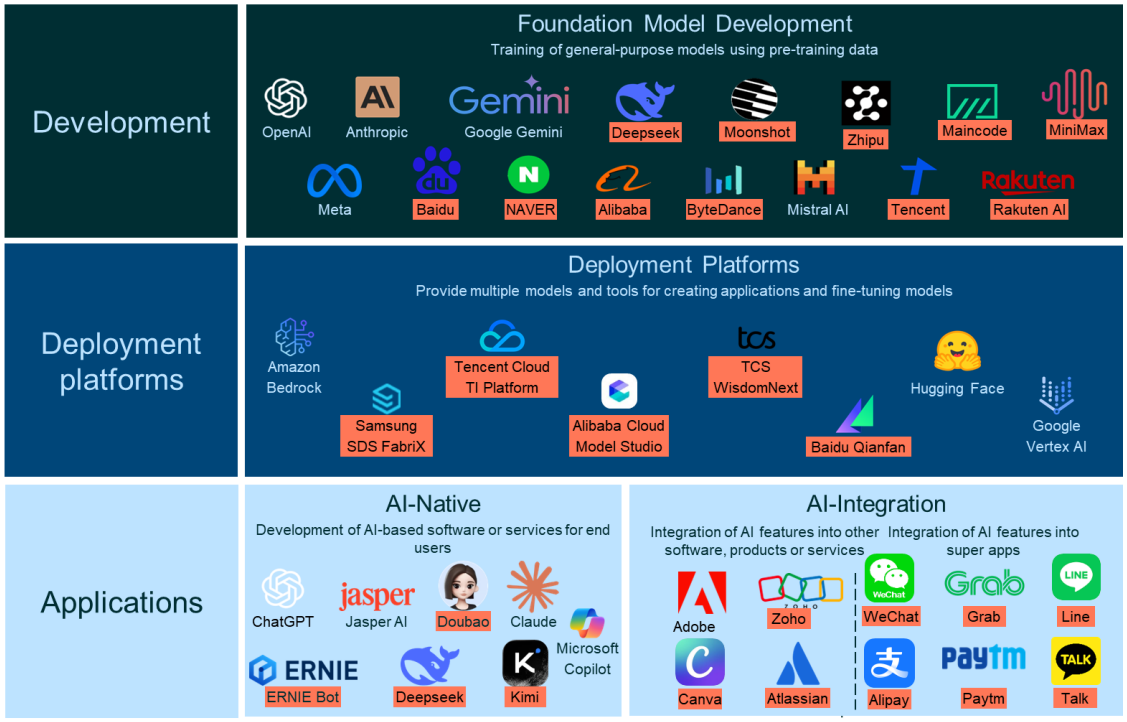
parameter model that outperforms models 10,000 times larger on specific reasoning benchmarks with a claimed training cost of less than USD 500. These lighter models reduce the infrastructure investment and computing power required to deploy GenAI features, making deployment accessible to a wider range of organisations.

1.3 Observed patterns are consistent with effective competition

Observed patterns across APAC support the view that competition is currently working effectively at the FM, deployment platform and application layers, speaking against the foreclosure and concentration concerns that competition authorities have identified as prospective risks. Our assessment instead suggests that multiple market-level mechanisms (open-source availability, orchestration tools, intermediaries, regulatory diversity and language-specific specialisation) reduce tipping risks and constrain concentration.

Consistent with the features discussed above, there are a large variety of suppliers operating at each level of the industry in APAC. This includes global and APAC-based players.

Figure 2: Selected key players at the development and deployment levels of the AI value chain in APAC



Source: *RBB analysis based on desk research of publicly available sources.*
Note: = *APAC-based player.*

1.4 Conclusion

Our assessment indicates that competition in the GenAI sector is currently working effectively, with vibrant competition among suppliers at the foundation model, deployment platform and application layers of the supply chain. We find no evidence of concerning levels of concentration based on current market evidence. Moreover, we observe a number of market features which will tend to promote competition, rather than consolidation.

In this context regulators should, in our view, think carefully before intervening in this nascent sector. At present, there is no evidence of adverse consumer outcomes resulting from limited competition. Moreover, given the rapid pace at which the technology is evolving, there are risks associated with premature intervention, which could disrupt effective competitive processes or chill innovation during critical development phases.

2 GenAI competitive landscape in the Asia-Pacific region

2.1 Introduction

This section examines the GenAI landscape in the APAC region. We first provide a short overview of the GenAI value chain. Next, we discuss important background trends that impact the evolution of the GenAI landscape in the region, including the regulatory environment, the mobile-first digital ecosystem, the prevalence of super apps, GenAI investment and the use of partnership structures.

We then examine recent developments at each level of the value chain that reflect the unique competitive forces at play in the region. Firms are moving to open-source, hybrid systems to stay flexible in their choice of supplier. We also describe how semiconductor export controls have led to rapid developments of GenAI models in the mature Chinese market, while domestic model development in other countries focuses on specialised local language tasks.

2.2 Overview of the GenAI value chain

In this paper we analyse three key levels of the GenAI value chain: (i) FM development, (ii) model deployment platforms, and (iii) applications.[34]

The **FM development** level builds the core technology on which GenAI relies. GenAI models are pre-trained on large quantities of data, creating FMs that have general capabilities in understanding language patterns or identifying objects in pictures or videos. These models can cover a wide range of different modalities including text generation, text-to-image, video generation, audio and speech, as well as multimodal models that span these areas. These models, such as OpenAI's suite of GPT models, Anthropic's Claude, Google's Gemini or Baidu's Ernie are then used by developers to drive GenAI capabilities in applications and services.

Model deployment platforms serve as market intermediaries between FM developers and GenAI deployers. They reduce the technical and commercial barriers to adoption by providing standardised interfaces to access models, tools for integration and customisation, and infrastructure for deployment. Without these platforms, deployers would need to negotiate separate relationships with each model provider and build bespoke infrastructure for each integration.

These intermediaries differ in the breadth of services they provide. At the most basic level, open-source repositories like Hugging Face and GitHub provide model access, documentation, and developer tools. Cloud-based platforms like Google Vertex AI, Amazon Bedrock, Alibaba Cloud Model Studio, and Baidu Qianfan add managed infrastructure, model gardens with multiple providers, and fine-tuning capabilities.[35] At the other end of the spectrum, service integrators like TCS provide fully managed solutions with on-site technical expertise and customised implementations.

At the **applications level**, businesses develop new GenAI applications or integrate GenAI models into existing products and services. This layer is broad, encompassing both specialist AI companies building GenAI products and the much larger universe of organisations integrating GenAI features into their

---

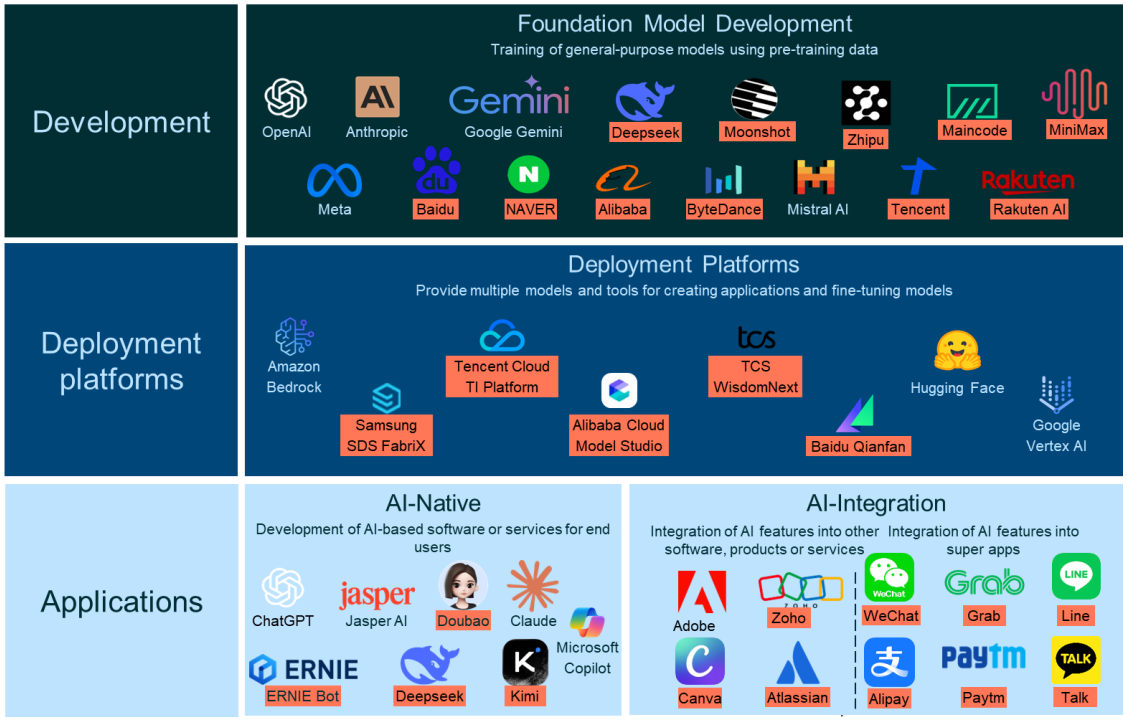
[34] A wider discussion could also encompass a further "infrastructure" level that provides inputs for FM development. These inputs include computing hardware (for example, GPUs, networking, memory, and storage for model design and training), cloud service providers (supply access to computing resources at usage-based pricing), training data, and technical expertise. For the purposes of this paper, we do not cover the infrastructure layer in detail.

[35] Following general pre-training, foundation models can be fine-tuned to improve performance. One aspect of this involves using reinforcement or instruction-based learning to teach the model how to follow instructions and provide output that aligns with human preferences. Another aspect involves adapting the model for specialised tasks by feeding it accurate, labelled data from specific domains. For example, this could include problems and answers to coding problems or annotated local language data. This fine-tuning can be conducted by model developers or application deployers.

operations or their products. Common use cases include AI chatbots (such as ChatGPT, ERNIE Bot, Clova X and Gemini), productivity assistants (such as Notion AI and Jasper), code-writing tools (such as GitHub Copilot and Codeium), and customer service virtual agents (such as Salesforce Einstein GPT, Freshworks Freddy and IBM watsonx Assistant). In APAC, super-apps like Grab, WeChat, Line, and Paytm have integrated GenAI features across their ecosystems. More broadly, enterprises across virtually every sector are deploying GenAI to enhance internal workflows, customer interactions, and product offerings.

Figure 3 below provides an overview of this value chain and the key players at each level. Firms can participate at multiple layers of the value chain and their presence can differ across countries or regions within APAC, depending on regulatory and commercial conditions.

Figure 3: Selected key players at the development and deployment levels of the AI value chain in APAC



Source: RBB analysis based on desk research of publicly available sources.

In the following sections, we discuss the general landscape and recent developments in APAC at each level of the value chain.

2.3 Regional market features

2.3.1 Regulatory environment

Regulatory approaches to AI and data governance vary dramatically across APAC. These regulations fundamentally affect the ability of certain players to compete in national markets, resulting in significant variations in the firms present nationally across the region.

- **Mainland China** maintains a restrictive environment for international competitors with a regulatory framework that promotes national models and content controls. The Cyberspace Administration of China requires GenAI FMs offering public services to obtain approval, with 346 GenAI services

registered as of April 2025.[36] Western developers such as OpenAI and Anthropic have not released their chat interfaces (ChatGPT and Claude) in China. In response to local developers using VPNs or accounts linked to foreign subsidiaries to access APIs for these models, both OpenAI and Anthropic restricted API access to Chinese and Hong Kong customers in 2024, citing compliance and security considerations (discussed further in the case study in section 5.2.1).[37] This promoted the adoption of models developed by national technology leaders such Alibaba's Qwen, Tencent's Hunyuan and Baidu's ERNIE. Despite the compliance requirements created by the registration process, the number of successful domestic registrations does not suggest this has been a barrier to domestic model and application development. China also maintains strict data residency requirements and mandates Critical Information Infrastructure Operators to store personal data, and important data with implications for economic and national security, within China.[38] This can require foreign cloud companies to access the market through local partners (for example, Microsoft Azure operates through 21Vianet as a separate entity in China).[39]

- In **Japan**, the regulatory environment is permissive. Japan's AI Promotion Act enacted in May 2025 establishes voluntary guidelines rather than mandatory requirements, reflecting a historical preference for self-regulation.[40] Japan's Act on the Protection of Personal Information (**APPI**) restricts personal data transfer to foreign countries unless (i) explicit consent is given, (ii) equivalent protection measures are undertaken by the data recipient, or (iii) the recipient is located in the EU or UK, regulatory regimes which the Japanese government deems offer similar protections to APPI.[41]

- **South Korea** implements a risk-based approach through its AI Framework Act (passed 2024, effective January 2026), distinguishing high-impact AI applications requiring assessment from other uses with lighter requirements.[42] The framework requires foreign AI providers meeting certain thresholds to appoint domestic representatives for compliance. South Korea's Personal Information Protection Act (**PIPA**) maintains stricter rules than Japan's APPI on user consent for international data transfers, imposing additional restrictions on data handlers, including security certifications and advance notice requirements for international server transfers, even if operated by the same provider.[43] Both Meta and SK Telecom have recently faced substantial penalties for data protection violations (involving fines of USD 97 million and USD 15.7 million, respectively).[44]

- **India's** AI Governance Guidelines promote the role of voluntary frameworks to tackle emerging risks. The guidelines emphasise the application of existing legislation to AI contexts, with a review of a small number of areas where updates may be required such as copyright law, digital platform classification and liability and content authentication.[45] The Digital Personal Data Protection Act (**DPDPA**) has received legislative approval, but some clauses have yet to be notified and come into full effect. The

---

36 These GenAI services are public-facing applications which fall under the scope of Article 2 of the Interim Measures for the Administration of Generative Artificial Intelligence Services. Consequently, downstream chatbots of Chinese FMs are subject to this regulation.
TechNode, 'China approves 346 generative AI services under national registration scheme', 9 April 2025.
Cyberspace Administration of China, 'Interim Measures for the Administration of Generative Artificial Intelligence Services', 13 July 2023.

37 Global Times, 'OpenAI's cut of China's access to its API service would 'push domestic developers to catch up vigorously'', 10 July 2024.
Anthropic, 'Updating restrictions of sales to unsupported regions', 5 September 2025.

38 Chambers and Partners, 'Cloud Computing 2025 – China', accessed November 2025.

39 Microsoft, 'Microsoft Azure operated by 21Vianet', accessed December 2025.

40 DLA Piper, 'Understanding AI Regulations in Japan', 28 October 2024.
International Bar Association, 'Japan's emerging AI framework for responsible AI', 16 July 2025.

41 Government of Japan, 'Act on the Protection of Personal Information', 30 May 2003, Article 28.
EU-Japan Centre for Industrial Cooperation, 'EU and Japan conclude first review of their bilateral mutual adequacy arrangement', June 2023.
UK government, 'UK-Japan Digital Partnership: Joint statement (January 2025)', 23 January 2025.

42 Legal500, 'A New Era for AI: Republic of Korea Takes a Bold Step with AI Regulation', 14 January 2025.

43 LexMundi, 'Global Data Privacy Guide - Korea, Republic of', accessed November 2025.
Government of South Korea, 'Act on the Development of Cloud Computing and Protection of its Users', 11 January 2022.

44 Reuters, 'South Korea agency fines SK Telecom $97 million over major data leak', 28 August 2025.
Reuters, 'South Korea fines about $15 mln over collection of user data', 6 November 2024.

45 Government of India Ministry of Electronics and Information Technology, 'India AI Governance Guidelines', 5 November 2025, pp 19-24.

act previously contained draft clauses requiring certain data types to be stored within India, but recent revisions remove this provision and instead give the government power to prevent personal data processing outside India in certain cases.[46] In addition, data handlers must provide detailed descriptions to data owners covering the type of data held and its usage.[47] The uncertainty over these compliance requirements has factored into organisations' IT strategies: 70% of Indian firms perceive regulatory compliance as a key risk of AI adoption.[48] This is significantly higher than the 42-43% of firms that voiced similar concerns in the US and France.[49]

The rest of APAC currently provides voluntary frameworks and guidance on AI governance and ethics, with most differentiation coming from the variability in data governance laws. **Hong Kong** maintains a separate regulatory environment from mainland China's registration system, operating under frameworks inherited from British administration and adapted through local legislation. It maintains a more permissive data regulation environment with draft legislation allowing data transfer across borders, with the provision that equivalent protections are provided as in Hong Kong.[50] **Malaysia** and **Singapore** likewise maintain relatively open approaches to personal data transfers under each country's Personal Data Protection Act (**PDPA**), with clauses allowing cross-border transfer to countries with similar levels of protection.[51]

Indonesia's Government Regulation 71/2019 and **Vietnam's** Decree 53/2022 require local data storage while remaining less restrictive than China by allowing cross-border transfers under specified conditions.[52] Given their alignment with Europe and the US, **Australia** and **New Zealand** remain the most open countries in the region, with local regulations prescribing that data handlers make reasonable efforts to ensure any international third-parties comply with relevant legislation.[53]

To help tackle differing data regulations between countries, the Association of Southeast Asian Nations (**ASEAN**) Model Contractual Clauses are available for use in agreements regarding the transfer of data between member states. In addition, the Asia-Pacific Economic Cooperation (**APEC**) Cross-Border Privacy Rules System facilitate cross-border data flows among APEC countries. However, these frameworks do not address all aspects of data compliance requirements and do not currently include all APAC countries.

This regulatory fragmentation increases compliance costs and affects business decisions. Unlike the EU's harmonised framework, APAC companies cannot deploy uniform strategies across the region. Geography compounds this: borders can prevent the efficient placement of key computing infrastructure and data storage. However, these same requirements drive organisations toward hybrid cloud architectures. Data sovereignty laws in Indonesia, Vietnam, Singapore and China require local data storage while permitting cross-border transfers under specified conditions. Organisations comply by maintaining relationships with multiple cloud providers, both local and international. This regulatory-driven multicloud adoption has a pro-competitive effect for GenAI: organisations already operating across multiple cloud platforms can access foundation models through multiple pathways, reducing dependence on any single model provider. That said, navigating diverse data residency laws remains more onerous for smaller entrants, who may need to partner with sophisticated providers that hold compliance capabilities across multiple jurisdictions.

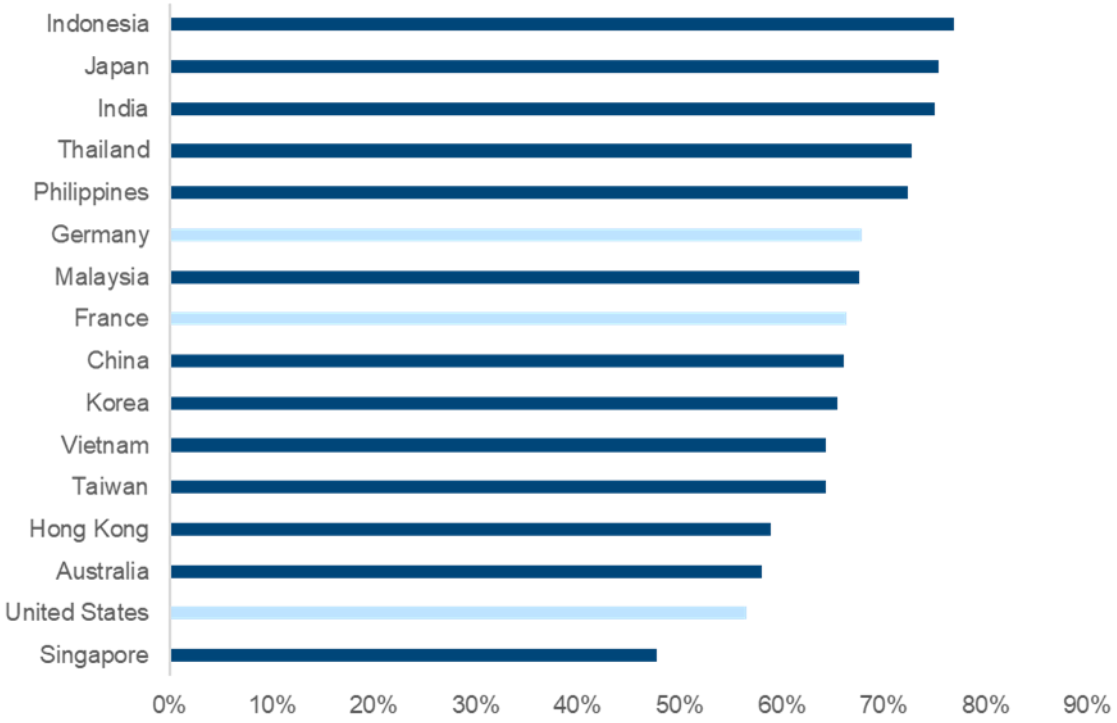
---

46 Mondaq, 'India's Digital Data Protection Act 2023: Key Insights and Comparisons With GDPR', 7 March 2025.
Tech Policy Press, 'Data Localization: India's Tryst with Data Sovereignty', 24 January 2025.
47 Ibid.
48 McKinsey, 'Open source technology in the age of AI', 22 April 2025, Exhibit 11.
49 Ibid.
50 CGJ HKCGI, 'Cross-border transfers of data', 13 September 2016.
51 Baker McKenzie, 'Malaysia: Personal Data Protection (Amendment) Act 2024 to come into force', 27 December 2024.
Singapore Statutes Online, 'Personal Data Protection Act 2012', accessed December 2025, part 6.
LexMundi, 'Global Data Privacy Guide - Singapore', accessed November 2025.
52 Global Compliance News, 'Indonesia: New Regulation on Electronic Systems and Transactions', 7 November 2019.
DLA Piper, 'Vietnam cybersecurity regulations for data storage', 18 October 2022.
53 Government of Australia, 'Privacy Act 1988', 1 February 2025, subclause 8.1.
Government of New Zealand, 'Privacy Act 2020', 30 June 2020, Section 22, Information Privacy Principle 12.

2.3.2 Mobile-first distribution

The majority of internet traffic in APAC countries originates from mobile devices as seen in Figure 4 below. India, Indonesia, China, Vietnam, the Philippines, Vietnam and Japan are among the top 15 mobile app download markets globally.[54] Consumers also spend significant time on mobile apps. Users in Indonesia, Singapore and South Korea spent five hours or more per day on mobile apps as of 2023.[55]

Figure 4: Mobile device traffic share by country, 2025



Source: RBB analysis based on Similarweb data for September 2025 (all countries bar China) and Statcounter data for June 2025 (China only).

As a result, mobile devices serve as primary access points for GenAI app developers across much of the region. GenAI app downloads in China rose to 133 million in 2024, up from 38.3 million in 2023.[56] Similarly, India hit 177 million downloads compared to 75.2 million in 2023.[57] As of 2024, APAC accounted for 55% of the global market for AI-in-mobile applications (defined in the study as apps that integrate machine learning and data analysis technologies).[58]

This mobile-first distribution also shapes interface design, leading developers to prioritise user interfaces suited to smartphone use. Voice and image inputs are particularly prominent across mobile devices.[59] Due to the mobile-first nature of APAC, image and voice inputs are more popular in this region than the rest of the world. For example, over 60% of consumers in Indonesia, China and India used voice on smartphones in 2022, compared to 42% globally.[60] More than one-third of Google Search queries in India are voice-based, compared to 5% in developed markets.[61] APAC is also projected to be the fastest

54 Sensor Tower, 'State of Mobile 2025', accessed November 2025, p10.
55 RipenApps, 'Mobile App Industry Statistics 2023: Trends and Insights You Shouldn't Ignore', 16 August 2023, Figure: Average Daily Time Spent on Apps by Country (hr).
56 Sensor Tower, 'State of Mobile 2025', accessed November 2025, p 21.
57 Ibid.
58 Markets.us, 'AI in Mobile Apps Market', March 2025.
59 Mobile devices were also the primary mode for image-based searches as of 2018, with 53% of image-based searches originating on mobile compared to 37% on laptop or desktop. DSI Spotlight, 'Visual Search Spotlight', October 2018.
60 This compares to 20-40% of consumers in developed markets such as Great Britain, France and the US. YouGov, 'Global: On which devices are consumers using voice commands?', 6 September 2022.
61 Ernst & Young, 'The AIdea of India 2025: How much productivity can GenAI unlock in India?', 14 January 2025, p 25.

growing region in image search due to its "*expanding digital commerce, rising mobile internet usage, and tech-savvy consumers*".[62]

Most GenAI applications today send queries to cloud servers for processing, then return results to the user's device. However, APAC's mobile-first ecosystem is driving interest in models small enough to run directly on smartphones or other local hardware. On-device processing offers faster response times, lower costs per query, and keeps user data on the device rather than transmitting it to external servers. These advantages are particularly valuable in mobile-centric markets with cost-sensitive users and varied data privacy requirements. As a result, APAC represents an important market for developers of compact, efficient models designed for on-device use (see section 2.4.1.6).

2.3.3 Super apps

In tandem with the mobile-first development trajectory, the APAC region has witnessed the rise of super app ecosystems that bundle multiple functions into single platforms. Super apps are an important route to access services and customers, resulting in incentives for developers to gain entry to the ecosystem through partnerships and distribution agreements. This has given rise to suites of "mini apps" which sit inside the main super app. These mini apps can range from shopping platforms and mobile gaming, to streaming services, grocery delivery and ride-hailing. These mini apps are popular with consumers as shown by WeChat mini apps achieving 400-450 million daily active users by 2020-2021.[63] As a result, some application developers increasingly operate primarily through super app ecosystems rather than maintaining independent platforms or websites.[64]

WeChat and Alipay are the largest super apps in the region. WeChat boasts over 1.4 billion monthly active users and offers a catalogue of over 4.3 million mini apps.[65] Originally launched as a messaging app in 2011, WeChat added social media features such as moments feeds in 2012. Payment capabilities followed soon after, including the popular Red Envelope function which allows users to send money to friends, as is customary in East Asian cultures for major celebrations, and spurred WeChat's entry into mobile payment systems.[66]

Alipay has almost identical reach, with 1.4 billion monthly active users.[67] Given its origin as a payment service to accompany the Alibaba e-commerce platform, Alipay maintains a slightly stronger position in mobile payments than WeChat, and is active in consumer-to-consumer marketplaces (Taobao, owned by Alipay's parent company, Alibaba) and food delivery (Ele.me, also owned by Alibaba) through a similar mini app marketplace.[68]

The open nature of these two ecosystems is demonstrated by cross-integration of capabilities between competing platforms. Alipay has been accepted as a payment method within Grab since 2016.[69] The partnership was expanded in 2025 through Alipay+ Voyager to allow users to book Grab rides directly

---

62 Data Bridge Market Research, 'Global Visual Search Market Size, Share, and Trends Analysis Report – Industry Overview and Forecast to 2032', May 2021.

63 Dauxe Consulting, 'Leveraging WeChat mini programs for enhanced brand engagement and independence', 18 March 2023.
Statista, 'Number of daily active users of WeChat Mini Program from 2017 to 2021', accessed November 2025.

64 Multiple businesses state that they no longer need to launch a standalone app in China. A WeChat mini app is sufficient.
AppInChina, 'WeChat: An essential tool for driving business growth in China', 10 May 2022
TMO Group, 'WeChat Mini-Programmes: 4 Successful Cases from Foreign Brands', 22 August 2024.

65 This figure includes the users for Weixin (China) and WeChat (international users). Tencent, '2025 First Quarter Results', 14 May 2025, p 2.
GMA, 'WeChat Statistics and User Trends for China in 2025', 5 August 2025.
Charlesworth, 'WeChat Essentials: Understanding WeChat Mini-Programmes', 8 April 2024.

66 Jiang, X. N., 'Analysis of WeChat Pay Based on Technology Acceptance Model', Advances in Economics, Business and Management Research, Vol. 215, 29 April 2022, p 669.

67 CoinLaw, 'Alipay Statistics 2025: User Adoption, Transaction Volumes, and Technological Innovations', 16 June 2025.

68 CoinLaw, 'Alipay vs. WeChat Pay Statistics 2025: Market Share, Innovation & Digital Yuan Impact', 3 August 2025.
Alibaba, 'Taobao', accessed December 2025.
Alibaba, 'Ele.me', accessed December 2025.

69 Grab, 'Grab and Alipay Offer Chinese Travellers an Easier Way to Pay for Rides in Singapore and Thailand', 21 June 2016.

from within the Alipay wallet across eight Southeast Asian countries.[70] More broadly, the Alipay+ network connects dozens of Asian e-wallets (including GrabPay, Touch 'n Go, Kakao Pay, and LINE Pay) enabling merchants to accept payments from multiple wallet providers.[71] Nearly 80% of enterprises that deploy mini apps do so on both WeChat and Alipay, with the platforms' similar technical frameworks enabling relatively straightforward conversion between them.[72] This highlights each super app's focus on providing consumers with best-in-class services.

The success of super apps is seen across the APAC region, with super apps gaining traction in India, Japan and Southeast Asia. Key APAC super apps and their functions are shown in Table 1.

[70] Grab, 'Alipay+ and Grab Make Ride Hailing Services Available to Global Digital Wallet Users via Alipay+ Voyager', 16 September 2025.
TechCrunch, 'Grab adds Alipay support to its taxi on-demand service in Southeast Asia', 20 June 2016.

[71] Alipay+, 'Alipay+ Mobile Payment Provider List', accessed December 2025.

[72] China Marketing Corp, 'WeChat Mini Programs', accessed November 2025.
Medium, 'We've experimented with Alipay Mini Programs ', 9 October 2017.
Web.dev, 'Mini app markup, styling, scripting, and updating', accessed November 2025.
Note, Alipay and WeChat mini apps are not fully interoperable. Rather, separate versions of popular mini apps are available on each platform's mini app marketplace.

Table 1: Summary of key APAC super apps

	WeChat	Alipay	paytm	LINE	Grab	TALK	gojek
Main country/region	China	China	India	Japan and Southeast Asia	Southeast Asia	South Korea	Indonesia
Monthly active users	1.4 billion[73]	1.4 billion[74]	270 million[75]	224 million[76]	46 million[77]	54 million[78]	38 million[79]
Founded	2011	2004	2010	2011	2012	2010	2010
Original use	Messaging	Digital wallet	Mobile payment plans	Messaging	Ride hailing	Messaging	Ride hailing
Mini app ecosystem	●	●	●	●	●[80]	●	●
Messaging services	●	◐[81]	○[82]	●	○	●	●[83]
Social media	●	◐[84]	○	●	○	●	○
Payment services	●	●	●	◐[85]	●	●	●
Mobile banking/financial services	●	●	●	◐[86]	●	●	●
E-commerce	●	●	●	●	◐[87]	●	●
Food delivery	●	●	●	◐[88]	●	●	●
Ride hailing	●	●	●	◐[89]	●	●	●
Gaming	●	●	●	●	●	●	◐[90]

Source: *RBB summary based on desk research of publicly available sources.*
Note: *● – Full functionality (core feature or mini apps); ◐ - Partial features or limited geographic scope; ○ – Not included.*

Super app ecosystems create different distribution dynamics to Western patterns where consumers tend to maintain separate relationships with messaging apps, payment providers and service platforms. The

[73] This figure includes the users for Weixin (China) and WeChat (international users). Tencent, '2025 First Quarter Results', 14 May 2025, p 2.
[74] CoinLaw, 'Alipay Statistics 2025: User Adoption, Transaction Volumes, and Technological Innovations', 16 June 2025.
[75] CoinLaw, 'Paytm Statistics 2025: Financial Performance and User Engagement Insights', 17 June 2025.
[76] Digital Marketing for Asia, 'Why is LINE the most popular social media app in Japan?', accessed December 2025.
[77] This figure is based on monthly transacting users. Grab, 'Grab Reports Second Quarter 2025 results', 31 July 2025.
[78] Market.biz, 'Kakaotalk Statistics and Facts', 20 August 2025.
[79] Techwire Asia, 'Gojek sees profitability ahead after decade of rapid growth', 16 November 2020.
[80] Grab, 'Grab launches third-party Partner Apps within Grab app, offering more everyday services for everyday needs', 27 October 2025.
[81] Jiemian Global, 'Alipay rolls out voice call feature with real-name display', 12 May 2025.
[82] Entrackr, 'Paytm kills in-app chat features from "Inbox"', 2 December 2018.
[83] Gojek, 'Chat', accessed November 2025.
[84] Jiemian Global, 'Alipay rolls out voice call feature with real-name display', 12 May 2025.
[85] The LINE Pay service was terminated in Japan as of 30 April 2025 as part of an agreement to move LINE Pay users onto PayPay digital wallets which is owned by the same ultimate parent company, SoftBank. LINE Pay is still operational in other markets such as Thailand and Taiwan.
LY Corporation, 'Termination of LINE Pay Service in Japan', 13 June 2024.
[86] LINE explored a potential joint venture to offer a "smartphone bank" with Mizuho Financial Group in Japan, however, the project was halted in 2023. Financial products and banking service are offered through the LINE super app in other jurisdictions such as Thailand.
LINE Corporation, 'LINE and Mizuho Financial Group Halt Project for New Bank', 30 March 2023.
Fintech Futures, 'Line launches messenger-based banking platform in Thailand', 22 October 2020.
[87] Grab offers GrabMart, their online grocery store, in major cities in all eight countries in which the Grab app is available.
[88] Food delivery functions are available in Thailand, but not Japan and Taiwan.
Bangkok Post, 'Company to build on Line Man, fintech', 2 July 2019.
[89] The Line app does not incorporate ride hailing services in Japan following discontinuation in 2018, but is available in Thailand.
McKinsey, 'Rebooting Japan's mobility market - Discussion paper', November 2018, p 7.
LINE Corporation, 'LINE Launches "LINE TAXI", A Taxi Calling Service in Bangkok', 2 April 2018.

open nature of these super app ecosystems allows internal and third-party developers to embed new features such as GenAI technologies into existing user interfaces with immediate distribution to sizeable user bases. Developers can also take advantage of the linkages between apps and centralised data collection to gain insights into the preferences of specific users and target content accordingly.

2.3.4 GenAI investment

The GenAI investment landscape in APAC differs from that seen in the EU and North America as private venture capital (**VC**) investment plays a less prominent role than government investment in the region. In addition, we see large cloud infrastructure providers commit significant funds to extending their public cloud networks across the region in part to service increasing AI workloads in the region, expanding the infrastructure available to local deployers.

Investments by established public cloud providers relate to both developed and developing economies in the region. AWS, Microsoft, Google and Oracle have collectively pledged nearly USD 27 billion in additional data centre development initiatives in Japan since 2024.[91] Alibaba Cloud and Tencent Cloud also have expansion plans with the launch of new data centres and increasing regional availability.[92] Cloud providers have also recognised Malaysia as a growing market with a cloud-friendly regulatory environment and strategic geographic location. AWS (USD 6.2 billion), Microsoft (USD 2.2 billion), Google (USD 2 billion), and Oracle (USD 6.5 billion) have all pledged significant sums for national infrastructure development in the country.[93]

Government investment targets all levels of the value chain, often mirroring current domestic strengths or perceived weaknesses, as APAC countries target sovereign and self-sufficient GenAI industries.

Table 2: Summary of public sector GenAI funding in APAC countries

Country	GenAI funding
China	Projected AI capital expenditure for 2025 ranges between USD 84-98 billion. Analysts project most of this investment will come from government funding (c. 57%) with large tech firms accounting for c. 24%.[94] Private VC funding in China is significantly lower at USD 140-400 million per year, compared to over USD 50 billion in the US in the first half of 2025 alone.[95]
Japan	Public sector investment of JPY10 trillion (USD 64.2 billion) announced for AI and semiconductors through 2030.[96, 97] This includes JPY 1.05 trillion (USD 6.74 billion) for next-generation chip development and JPY 471.4 billion (USD 3.03 billion) to support domestic production of advanced chips.[98]

90 Gojek launched GoGames in 2019, and updates to the service were documented on its internal blog as recently as 2024. However, GoGames is no longer listed on the website as an active feature, and it is unclear whether it is still operational.
KrASIA, 'Gojek launches GoGames, bringing full ecosystem to Indonesia's mobile game industry', 9 September 2019.
Gojek, 'Gogames', accessed November 2025.
Gojek, 'Help', accessed November 2025.

91 CRN, 'AWS Pours $15B Into Japan As AI Race With Microsoft, Google Continues', 19 January 2024.
Microsoft, 'Accelerating Japan's growth with AI', 27 March 2025.
Google Cloud, 'Investing $1 billion in digital connectivity to Japan', 11 April 2024.
Oracle, 'Oracle to Invest More Than $8 Billion in Cloud Computing and AI in Japan', 17 April 2024.

92 Datacenter Dynamics, 'Alibaba Cloud to launch data centers in eight locations in coming year', 25 September 2025.
Tencent Cloud, Tencent Cloud Japan to Expand Support for Japan's Digital Economy Setting up Osaka as new Cloud Region with its 3rd Data Center in Japan', 16 April 2025.

93 Microsoft, 'Microsoft announces US$2.2 billion investment to fuel Malaysia's cloud and AI transformation', 2 May 2024.
Amazon, 'AWS launches Malaysia's first cloud infrastructure region', 22 August 2024.
Google Cloud, 'Advancing Malaysia Together: Google Announces US$2 Billion Investment in Malaysia, Including First Google Data Center and Google Cloud Region', 30 May 2024.
Oracle, 'Oracle to Invest More Than US$6.5 Billion in AI and Cloud Computing in Malaysia', 2 October 2024.

94 TechWire Asia, 'China to deploy $98bn in AI investment this year amid US tech rivalry', 26 June 2025.

95 GlobalData, 'GenAI VC funding in early 2025 highlights widening gap between US and China, finds GlobalData', 27 May 2025.

96 Reuters, 'Japan unveils $65 bln plan to aid domestic chip industry', 12 November 2024.

97 IMF Representative Rates for 1 December 2025 are used for currency conversions unless stated otherwise. See IMF.

98 Quantum Insider, 'Japan Boosts Semiconductor, Quantum R&D with Trillion-Yen Budget', 16 January 2025.

	The Ministry of Economy, Trade and Industry has also allocated up to JPY 72.5 billion (USD 465.3 million) to five national firms to develop cloud services for AI computing infrastructure.[99]
South Korea	Investment in technology sectors including AI reached KRW 3.6 trillion (USD 2.46 billion) in 2024, representing a 34% increase year-on-year.[100] The pursuit of sovereign AI has led to KRW 530 billion (USD 362 million) government investment in five domestic AI champions: LG AI Research, NC AI, Naver Cloud, SK Telecom and Upstage.[101] However, South Korea remains open to international firms with OpenAI opening a Seoul office in 2025 even as South Korea develops independent GenAI capabilities.[102]
India	The IndiaAI Mission allocates INR 103.72 billion (USD 1.16 billion) over five years toward computing infrastructure, datasets and AI capability development.[103] As part of this project, INR 9.9 billion (USD 110 million) has been allocated to the BharatGen project which aims to create multimodal models for 22 Indic languages.[104]
Singapore	AI Singapore initiative channels SGD 500 million (USD 386 million) into partnerships and infrastructure development, positioning the city-state as a regional AI hub.[105] The National Research Foundation has committed SGD 70 million (USD 54.0 million) to fund the development of the Meralion and Sealion FM families adapted to 11 Southeast Asian languages and cultures.[106]
Indonesia	Indonesia's USD 10 billion sovereign wealth fund, INA, is increasingly targeting AI initiatives and digital infrastructure. A partnership between INA and a Singapore-based asset management firm led to a joint investment in Indonesia's tech and wider AI ecosystem.[107]
Malaysia	Malaysia's government has launched the National AI Office with funding commitments of RM 10 million (USD 2.42 million). An additional RM 50 million (USD 12.1 million) has been pledged for AI education to help the country become a regional hub for AI development and talent.[108]
Taiwan	Public sector investments include NTD 10 billion (USD 318 million) for the Enhanced AI Startup Investment Programme and initiatives providing free GPU computing resources to AI startups.[109, 110]
Australia	Australia is an outlier in the region with a greater commitment to encouraging private sector investment in GenAI through mechanisms such as tax incentive schemes.[111] The Australian government has recognised the opportunities of GenAI and has allocated capital from the AUD 1 billion (USD 654 million) National Reconstruction fund.[112] Additionally, AUD 124.1 million (USD 81.2 million) has been earmarked for the government's AI Action Plan which includes the creation of a National Artificial Intelligence Centre, the support of AI and GenAI research and upskilling graduates, as well as piloting AI and GenAI use cases through public-private partnerships.[113]

Source: RBB analysis based on desk research of publicly available sources.

This investment activity supports the range of partnership structures and competitive entry observed across the region, discussed in the following section.

99 Ministry of Economy, Trade and Industry, 'Approval of Plans for Ensuring a Stable Supply of Cloud Programmes under the Economic Security promotion Act', 19 April 2024.
100 KoreaTechDesk, 'South Korea's Deep Tech Investment Hits Record High in 2024', 10 April 2025.
101 TechCrunch, 'How South Korea plans to best OpenAI, Google, others with homegrown AI', 27 September 2025.
102 OpenAI, 'AI in South Korea—OpenAI's Economic Blueprint', 23 October 2025.
103 India AI, 'Cabinet approves India AI mission at an outlay of Rs 10,372 crore', 12 March 2024.
104 India Ministry of Science and Technology, 'Union Minister Dr. Jitendra Singh launches 'Bharat Gen'', 2 June 2025.
BharatGen, 'India's Sovereign AI BharatGen Secures ₹988.6 Crore Under IndiaAI Mission', 18 September 2025.
105 Singapore Economic Development Board, 'Singapore's National AI Strategy: AI for the public good, for Singapore and the world', 4 December 2023.
106 Infocomm Media Development Authority, 'National Multimodal LLM Programme', accessed November 2025.
Sea-Lion.AI, 'Catalysing AI Innovation for Southeast Asia's languages', accessed December 2025.
107 Reuters, 'Indonesia Sovereign wealth fund INA targets data centres, AI in healthcare, renewables', 17 September 2025.
108 Malaysian Investment Development Authority, 'Budget 2025 to accelerate Malaysia's digitalisation, AI adoption', 20 October 2024.
109 Taiwan Ministry of Digital Affairs, 'Ministry of Digital Affairs Unveils Major Initiatives to Boost Taiwan's AI Industry and Global Competitiveness', 24 March 2025.
110 Exchange rate for New Taiwanese Dollar - 1 USD: 31.411 NTD. See Exchange-Rates.org
111 Australia Department of Industry, Science and Resources, 'Developing a National AI Capability Plan', 13 December 2024.
112 Ibid.
113 Australian Government, 'Digital Economy Strategy 2030', 11 May 2021, p 65.

2.3.5 Partnership structures

Private and public sector partnerships exist across the APAC GenAI space. These cover commercial activities such as model development and deployment, as well as broader policy objectives including data governance and national AI capability building.

Private sector partnerships meet a range of user needs. Partnerships between application deployers and model developers or model deployment platforms help less experienced adopters incorporate GenAI FMs via pre-packaged products, cutting deployment time. More experienced or technically competent deployers may partner with FM developers to test and refine model integrations, or work with platforms and integrators to develop tailored data pipelines or orchestration systems (see further discussion in section 4.3). Deployment platforms partner with FM developers and application developers to broaden their offerings and better meet customer needs.

Examples of private sector partnerships across the region include:

- As part of the AI Accelerator programme in Singapore, Google partnered with 84 organisations to develop an AI sandbox allowing engineers to create and test new GenAI products in a secure development environment. This sandbox included access to cloud infrastructure, Google's Vertex developer platform, multiple FMs and low-code developer tools.[114]
- Microsoft entered a partnership with Singaporean data governance company Straits Interactive, which involved incorporating Straits' Data Protection Officer Assistant into Azure's OpenAI service. This AI assistant is trained on ASEAN AI ethics and data protection legislation and can provide data governance advice accounting for internal company policies and best-practice guidelines.[115]
- Alibaba Cloud's Partner Rainforest Plan serves 50 AI technology partners through an AI Alliance Accelerator Programme. Partners such as Yell Group in Thailand use Alibaba's technology to develop new GenAI solutions for the creative media industry.[116]
- Huawei has over 400 partners, including AIS in Thailand, with whom it jointly develops GenAI solutions for wireless network management.[117]
- OpenAI has established a joint venture with Japan's Softbank to develop a pre-packaged, enterprise solution for corporate management tasks in Japan.[118]
- CBA has entered co-development partnerships with AWS, Anthropic, Microsoft and OpenAI as part of its plan to use GenAI to improve internal productivity and banking services in Australia (see section 5.2.6 for further discussion).

APAC governments also actively seek partnerships with other public bodies, research institutions and industry organisations in the GenAI space to develop sovereign AI capabilities (see section 2.3.4). These initiatives include:

- **Cross-border AI development sandboxes**, such as the recently launched initiative between regulators in Singapore, Malaysia, Thailand, and China, which provides a shared environment for governments to experiment with AI governance and test models across jurisdictions;[119]

---

114 EDB Singapore, 'AI Trailblazers Initiative: How companies are using GenAI in Singapore', 16 April 2024.
115 Microsoft, 'Straits Interactive to Advance Responsible AI, Ethics, and Data Privacy in Singapore and the ASEAN Region', 25 July 2023.
116 Alibaba Cloud, 'Alibaba Cloud revamps Global Partnership Ecosystem', 3 December 2024.
117 Huawei Cloud, 'Forging an AI Cloud Foundation: Huawei Cloud Accelerates Intelligence with APAC partners', 9 May 2025.
Huawei, 'AIS and Huawei launch RAN Intelligence Pioneers Programme to Expedite AN L4 Evolution', 9 October 2024.
118 SoftBank Group, 'The SoftBank Group and OpenAI Launch "SB OAI Japan" Joint Venture', 5 November 2025.
119 Microsoft, 'Asia Pacific's AI Leap: From Strategic Drive to Agentic Innovation', 6 November 2025.

- **Data acquisition partnerships for local language model development**, including Singapore's National Multimodal LLM Programme, where government agencies, research institutes and industry partners pool regional linguistic data to build models tailored to Southeast Asian languages; and India's AI Mission, which provides national supercomputing and data infrastructure to researchers, startups and industry to support training of locally relevant AI systems;[120]
- **Identification of potential GenAI use cases**, such as Australia's National AI Capability Plan, which engages industry, unions and civil society to identify areas where AI can deliver economic impact and where national capability may be required;[121]
- **Recruitment and training of AI talent**, as seen in South Korea's plan to invest KRW 300 billion (USD 205 million) to attract 400 domestic and international postdoctoral researchers and strengthen its AI research ecosystem;[122] and
- **Targeted government funding** (as shown in Table 2 in section 2.3.4).

These initiatives emphasise the important role governments in the APAC region play in shaping the trajectory of AI development and deployment.

2.4 Recent APAC developments in the GenAI value chain

2.4.1 FM development

2.4.1.1 Introduction

FM development and usage follow different trends across APAC countries, notably between China and other APAC countries.

China's closed ecosystem has led to the development of a wide range of general FMs which garner significant usage both domestically and internationally. Many of these models can be accessed through consumer-facing chatbot applications. ByteDance's Doubao model reaches the widest audience through its popular AI chatbot, with up to 172 million monthly active users, followed by DeepSeek with an estimated 145 million monthly active users.[123] Others, such as Kimi and ERNIE Bot have also amassed significant downloads and user bases. Alibaba's open-source Qwen models are popular among developer communities and have recorded 610 million downloads with over 170,000 derivative implementations created by developers, with Qwen2.5 ranking first globally in total downloads among all open-source FMs on Hugging Face.[124]

By contrast, Western foundation models feature prominently in the remaining APAC countries, given their smaller domestic markets and funding scales.

- India ranks as the second highest origin of ChatGPT website traffic (after the US), with Japan in fourth place.[125]
- Google Gemini shows strong usage in India, Japan and Indonesia.[126]

---

120 Infocomm Media Development Authority, 'National Multimodal LLM Programme', accessed November 2025.
India AI, 'Cabinet approves India AI mission at an outlay of Rs 10,372 crore', 12 March 2024.
AI4Bharat, 'Building AI for India! ', accessed December 2025.
121 Australia Department of Industry, Science and Resources, 'Developing a National AI Capability Plan', 13 December 2024.
122 Donga, 'South Korea launches global drive for AI talent', 16 June 2025.
123 Forbes, 'AI Mania Makes ByteDance Cofounder Zhang Yiming China's Richest Person', 9 March 2025.
Tech in Asia, 'Most Chinese AI apps losing users despite 700 million base: report', 30 October 2025.
124 Hugging Face, 'Model statistics of the 50 most downloaded entities on Hugging Face', 13 October 2025, Figure 4.
Alizila, 'Alibaba Recognized on Fortune's 2025 Change the World List for Open-Source AI', 25 September 2025.
125 Similarweb, 'chatgpt.com Website Analysis for October 2025', accessed November 2025.
126 DOIT Software, 'Google Gemini Statistics', 10 November 2025.

- Anthropic's Claude model counts India, Japan and South Korea among its top five user bases in terms of absolute size, while finding high usage per capita in Singapore, Australia and New Zealand.[127]

Model development outside of China focuses primarily on fine-tuning models for specialised use cases such as domestic language interpretation and generation, or tasks requiring strong working knowledge of local culture.

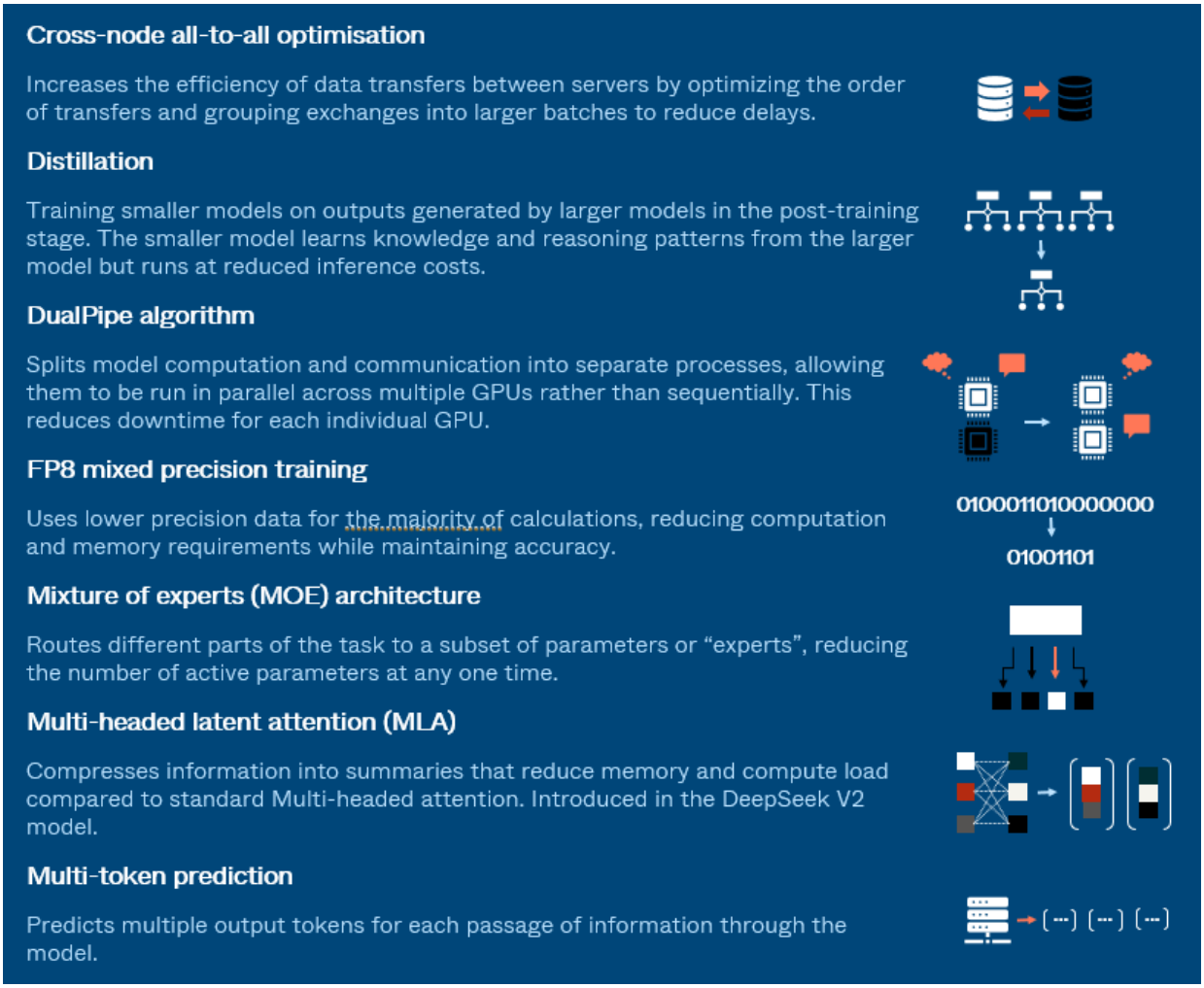
2.4.1.2 Advances in FM development

Restrictions on exports of advanced semiconductors to China have influenced competitive strategies and may have helped drive efficiency improvements. This limited access to high-end chips has pushed Chinese FM developers to design highly efficient models and training methods to reduce computing costs. These efficiency gains have enabled Chinese model providers to release competitive open-weight models and offer API services at much lower prices than many international alternatives.

The launch of DeepSeek-V3 in December 2024 and DeepSeek-R1 in January 2025 exemplify this innovation in model development. DeepSeek-V3 achieved comparable benchmark performance to GPT-4o while claiming training costs of only USD 5.58 million.[128] This represents a fraction of the USD 100 million or more spent on earlier frontier models such as OpenAI's GPT-4.[129] Both models were released as open-source, allowing developers worldwide to run them locally rather than relying on API access. The DeepSeek-V3 model uses novel approaches as outlined in Figure 5 below.

[127] Anthropic, 'Anthropic Economic Index report: Uneven geographic and enterprise AI adoptions', 16 September 2025, Figures 2.1 and 2.2.
[128] DeepSeek, 'DeepSeek-V3 Technical Report', 27 December 2024.
[129] Wall Street Journal, 'The Next Great Leap in AI Is Behind Schedule and Crazy Expensive', 20 December 2024.

Figure 5: DeepSeek-V3 novel approaches



Source: RBB summary based on DeepSeek, 'DeepSeek-V3 Technical Report', 27 December 2024.

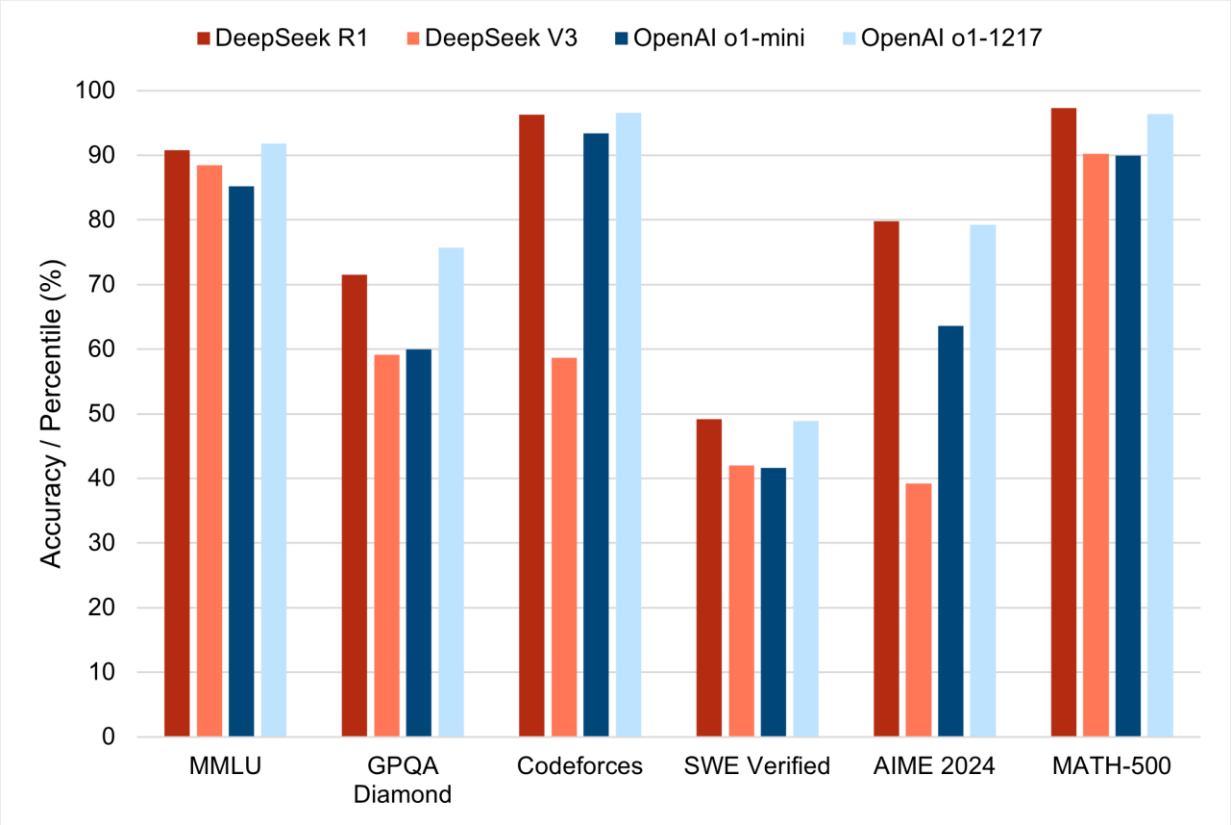
DeepSeek-R1 builds on the V3 model by refining the post-training process.[130] The standard post-training process involves two stages: (i) feeding the model labelled data to learn correct answers across a range of topics, and (ii) using reinforcement learning from human feedback (**RLHF**) or reinforcement learning from AI feedback (**RLAIF**) to learn human preferences by rewarding the model based on the palatability of its answers. DeepSeek-R1 was instead trained on a much smaller synthetic labelled dataset before engaging in reinforcement learning with automated reward signals based on verifiable answers in maths and code using a novel reward algorithm.[131] This significantly reduced the data and time requirements to develop the model. The overall post-training process of DeepSeek-R1 was estimated to have cost as little as USD 294,000.[132]

Despite this streamlined approach, DeepSeek-R1 matched or exceeded OpenAI's o1 reasoning model's performance on four out of six major benchmarks, with particularly strong performance on mathematics and coding tasks. The comparison of DeepSeek-R1 and DeepSeek-V3 with OpenAI o1 models across a range of benchmarks is shown in Figure 6 below. These benchmarks cover a wide array of topics including:

130 DeepSeek, 'DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning', 22 January 2025.
131 Ibid.
132 Reuters, 'China's DeepSeek says its hit AI model cost just $294,000 to train', 19 September 2025.

- general capabilities (Massive Multitask Language Understanding (**MMLU**) and Graduate-Level Google-Proof Q&A (**GPQA Diamond**));
- coding algorithms (Codeforces);
- engineering-oriented coding tasks (SWE Verified); and
- mathematics tasks (American Invitational Mathematics Examination (**AIME**) 2024 and MATH-500).

Figure 6: Comparison of DeepSeek versus OpenAI on various AI benchmarks



Source: RBB analysis based on DeepSeek, 'DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning', 22 January 2025, Table 4.

These training breakthroughs and cost reductions led DeepSeek to reduce API off-peak access prices by up to 75% in early 2025, with additional 50% price cuts for its new V3.2 model in September 2025.[133] These moves forced aggressive price responses from rivals, which we discuss in greater detail in the China landscape case study in section 5.1.

DeepSeek is not the only developer realising significant efficiency gains in model training and development. For example, Moonshot AI released Kimi K2 Thinking, which performs similarly to US models but purportedly cost only USD 4.6 million to train.[134]

In addition to these well-publicised efficiency gains in China, researchers across APAC are driving related innovations in model development. Researchers in South Korea are developing new quantisation methods that reduce memory usage and training costs by efficiently altering the numerical precision of model parameters for more efficient fine-tuning.[135] Japanese conglomerate Fujitsu has also contributed

[133] Reuters, 'DeepSeek cuts off-peak pricing for developers by up to 75%', 26 February 2025.
VentureBeat, 'DeepSeek's new V3.2-Exp model cuts API pricing in half to less than 3 cents per 1M input token', 29 September 2025.
[134] CNBC, 'Alibaba-backed Moonshot releases its second AI update in four months as China's AI race heats up', 6 November 2025.
[135] Jeon et al., 'L4Q: Parameter Efficient Quantization-Aware Fine-Tuning on Large Language Models', 2025.
Lee et al., 'QEFT: Quantization for Efficient Fine-Tuning of LLMs', 2024.

to this area with the development of a new quantisation optimisation algorithm that prevents errors from accumulating across model layers.[136] This algorithm led to reducing memory consumption by up to 94% while retaining 89% of model accuracy, enabling large models to run on edge devices such as smartphones and factory equipment rather than data centres.[137] Rakuten, a Japanese e-commerce and fintech firm with an extensive ecosystem, has developed a custom Japanese tokeniser algorithm which greatly reduces the cost of processing data for FM training, highlighting both the commitment to model efficiency and also the incentives to cater for local language requirements (see further discussion in section 2.4.1.4).[138]

These efficiency gains suggest a dynamic innovation environment in the APAC region, with developers increasingly prioritising cost-effective training approaches alongside raw performance. While export controls on advanced semiconductors have constrained hardware access for some Chinese developers, it is difficult to establish whether these restrictions have accelerated efficiency-focused innovation or whether such innovation would have occurred regardless. What is clear is that the constraints have not prevented Chinese developers from producing competitive models, and the resulting efficiency techniques are now available globally through open-weight releases.

2.4.1.3 Prevalence of open-source models in APAC

Open-source FMs are prevalent in the APAC region and play a central role in the GenAI ecosystem. As shown in the table below, a wide range of leading Chinese FMs have been released as open-source.

136 IT Business Today, 'Fujitsu enhances Takane with lighter AI technology', 8 September 2025.
137 Ibid.
138 Rakuten, 'Rakuten Releases High-Performance Open Large Language Models Optimized for the Japanese Language', 21 March 2024.

Table 3: Overview of prevalent Chinese FMs

Company	Key Model	Open-Source	Comment
deepseek	DeepSeek-R1[139]	●	
Alibaba	Qwen[140]	◐	Some Qwen variants are open-source, but certain models (for example, Qwen-2.5-Max) remain proprietary.
ZHIPU·AI	GLM[141]	●	
Baidu 百度	ERNIE[142]	●	ERNIE 4.5 became open-source in July 2025, but Baidu's earlier models were proprietary.
Tencent 腾讯	Hunyuan[143]	◐	Hunyuan has several open-source models (for example, Hunyuan-Large) with public code and weights, while other variants remain closed.
ByteDance	Doubao[144]	○	Core Doubao models are closed-source with no public weights or model code; access is API-only.
MINIMAX	MiniMax (M1 and M2)[145]	●	Flagship M2 is open and publicly released, though some other MiniMax models may be proprietary.
Moonshot AI	Kimi[146]	●	
百川智能 BAICHUAN AI	Baichuan[147]	●	
01.AI	Yi-Lightning[148]	●	
StepFun	Step3[149]	●	

Source: *RBB summary based on desk research of publicly available sources.*
Note: ● *– Yes;* ◐ *– Partial;* ○ *– No.*

In the ASEAN region, GenAI startups report using a mix of proprietary and open-source FMs. OpenAI leads the way with usage by 76% of respondents, followed by Meta's open-source Llama model (36%), open-source models accessed through Hugging Face (33%) and Anthropic (31%).[150] Indian startups have a strong preference for open-source with 76% of startups building applications mainly on open-source

139 Github, 'DeepSeek-R1', accessed November 2025.
DeepSeek, 'DeepSeek-R1 Release', 20 January 2025.
140 Github, 'Qwen', accessed November 2025.
CometAPI, 'Alibaba's Qwen: Is It Truly Open Source?', 20 April 2025.
141 Github, 'GLM-4', accessed November 2025.
Hugging Face, 'GLM-4.5', accessed November 2025.
Implicator.ai, 'GLM-4.6 puts receipts on the table: open weights, real coding runs, cheaper tokens', 1 October 2025.
142 Technode, 'Baidu open-sources ERNIE 4.5 series models, including multimodal MoE architecture', 1 July 2025.
Github, 'ERNIE', accessed November 2025.
Reuters, 'China's Baidu to make latest Ernie AI model open-source as competition heats up', 14 February 2025.
143 Github, 'Tencent-Hunyuan', accessed November 2025.
Tencent Hunyuan, 'Hunyuan Open-Source Models', accessed December 2025.
The Decoder, 'Tencent's Hunyuan-Large-Vision sets a new benchmark as China's leading multimodal model', 17 August 2025.
144 SCMP, 'China's most popular AI app gets facelift with ByteDance's Doubao 1.5', 22 January 2025.
ByteDance, 'Doubao-1.5-pro', 22 January 2025.
145 Asia Tech Lens, 'Meet MiniMax: The Chinese Tech Company Touted by Jensen Huang That's Headed for an IPO', 30 July 2025.
Github, 'MiniMax-M1', accessed November 2025.
Github, 'MiniMax-M2', accessed November 2025.
146 Hugging Face, 'Kimi-K2-Instruct-0905', accessed November 2025.
Github, 'Kimi-K2', accessed November 2025.
147 Github, 'Baichuan Intelligent Technology', accessed November 2025.
Hugging Face, 'Baichuan-7B', accessed November 2025.
148 Pandaily, '01.AI Releases New Flagship Model Yi-Lightning', 19 October 2024.
Github, 'Yi', accessed November 2025.
149 Github, 'Step3', accessed November 2025.
Hugging Face, 'Step3', accessed November 2025.
150 GenAI Fund, 'ASEAN GenAI Startup Report 2024', 15 September 2024, p 23.

models and 17% primarily on closed source models.[151] Even when surveying a broader spectrum of organisations, Indian organisations are more likely to regularly use open-source AI models (77%) than their peers in economies such as the UK (66%), US (62%) or France (56%).[152]

Government support reinforces open-source strategies. Facing US export controls on advanced semiconductors, Chinese authorities have made "independent and controllable" AI a key policy objective, and view domestic open-source ecosystems as critical for technology independence and as a way to accelerate industry progress.[153] Meta has pursued a different rationale for open-sourcing Llama: avoiding dependency on competitors' closed ecosystems, building a broad developer ecosystem around its FMs, and recognising that selling model access is not Meta's core business.[154]

Open-source models can be freely downloaded and used, allowing for wider dissemination of FMs within developer communities. The rapid adoption of open-source models is reflected in the scientific community, where DeepSeek tools were cited in 38% of all new AI research papers on Arxiv in Q1 2025.[155]

Developers who lack the resources to train an FM from scratch (which can cost millions of dollars) can instead build applications using existing open-source models at minimal cost. They can also fine-tune models for specific uses and offer them to other developers. As discussed in section 4.1, this lowers barriers to entry, increases the number of viable competitors, and lowers switching costs between model providers, supporting more diverse and dynamic competition.

2.4.1.4 Local language optimisation and multimodal requirements

APAC's linguistic diversity creates relative advantages for local model developers that allow them to compete with leading multilingual models from international developers in local languages.

The region encompasses languages with fundamentally different characteristics from Indo-European language families.

- Chinese languages use logographic writing systems with thousands of distinct characters.
- Japanese combines multiple writing systems (kanji, hiragana and katakana) and requires different encoding and processing approaches for each.
- Korean also employs a unique phonetic alphabet structure that requires high-quality language training data.
- Southeast Asian languages span multiple language families with distinct grammatical patterns. In particular, models used in Southeast Asian countries must handle code-switching between languages where users routinely mix English, Mandarin and local languages within single conversations.

Because many APAC markets are mobile-first, users frequently interact with services through voice, images as well as text, which in turn requires model developers to build multimodal models that can process these formats.

These linguistic and technical requirements mean that international developers cannot simply deploy English-optimised models in APAC market and expect them to perform as well as in their markets of origin.

---

151 Competition Commission of India, 'Market Study on Artificial Intelligence and Competition', September 2025, p 23.
152 McKinsey, 'Open source technology in the age of AI', 22 April 2025, p 6.
153 Merics, 'China's drive toward self-reliance in artificial intelligence: from chips to large language models', 22 July 2025.
RAND, 'Full Stack: China's Evolving Industrial Policy for AI', 26 June 2025.
Techwire Asia, 'The China open-source AI revolution that's rattling Silicon Valley', 24 July 2025.
154 Meta, 'Open Source AI is the Path Forward', 23 July 2024.
155 SQ Magazine, 'DeepSeek AI Statistics 2025', 7 October 2025.

This opens opportunities for domestic developers to produce fine-tuned models that better handle local languages, dialects, and multimodal inputs, that compete against global players and boost competition.

- Naver's HyperClova X is trained on 6,500 times more Korean data than GPT-4, reflecting Naver's privileged access to large-scale Korean-language content across its search engine, user-generated services and wider digital platform ecosystem, and the relative scarcity of Korean-language training data available to international developers.[156] HyperClova X demonstrates meaningfully superior performance on Korean language benchmarks against a number of competing models developed for Korean-language tasks or multilingual tasks (including Falcon, SOLAR and Llama).[157] However, new models such as OpenAI's o1, GPT-4o and Anthropic's Claude 3.5 Sonnet outperformed HyperClova X on the Korean College Scholastic Ability Test, indicating leading multilingual models from global developers may be closing the gap.[158]
- Baidu's ERNIE models, optimised for Chinese language nuances including semantic context and cultural references, maintain performance advantages in Chinese applications over international competitors.[159]
- Chinese-developed models consistently outperform international alternatives on Chinese language benchmarks, with models from Alibaba, ByteDance and DeepSeek leading major leaderboards.[160]
- In India, developers have launched multiple models targeting local language requirements, including Tech Mahindra's Project Indus, Sarvam 1, and SUTRA by TWO.ai.[161] India's 22 official languages mean models must support numerous languages to gain wide-ranging adoption. Sarvam 1, for instance, operates in ten Indic languages and claims to perform well compared to larger parameter Llama models and Google's similarly sized Gemma model.[162]

Cultural or local knowledge also offer routes to market for local developers. In many local contexts, models must understand cultural references, historical background, and social norms to provide useful outputs. For example, Korean models must navigate a complex honorific system where verb forms and vocabulary shift based on the relative status of speakers, a feature embedded in grammar that generic multilingual models often handle poorly. Similar dynamics apply even for English-language markets. In Australia, Maincode and Sovereign Australia AI have announced the country's first domestic FM development efforts.[163] These initiatives reflect demand for models trained on Australian data and deployed on local infrastructure, driven partly by data sovereignty considerations alongside expectations of improved performance for Australian queries.[164]

This recent progress in the APAC region and the significant support through government investment and partnerships (see sections 2.3.4 and 2.3.5) demonstrates that both the private and public sectors see a significant role for language-tailored FMs. Domestic firms are taking advantage of this opportunity to provide their own models.

However, local language advantages do not automatically lead to a single or a few providers. Multiple domestic developers are competing in each language market: China has numerous competing models,

156 The Korea Herald, 'Korea's AI challengers take on ChatGPT with own LLMs', 1 September 2025.
157 Naver Cloud, 'HyperClova X Technical Report', 13 April 2024, Figure 1.
158 GitHub, 'Korean SAT LLM Leaderboard', accessed November 2025.
159 Baidu Research, 'ERNIE Bot: Baidu's Knowledge-Enhanced Large Language Model Built on Full AI Stack Technology', 24 March 2023.
Baidu, 'ERNIE 4.5 Technical Report', 29 June 2025.
160 Hugging Face, 'Open Chinese LLM Leaderboard', accessed November 2025.
GitHub, 'Chinese LLM Benchmark', accessed November 2025.
161 Ernst & Young, 'The Aldea of India 2025: How much productivity can GenAI unlock in India?', 14 January 2025, p 30.
Nasscomm, 'India's Generative AI Startup Landscape 2024', 16 October 2024, p 39.
162 Sarvam, 'Sarvam 1: The first Indian language LLM', 24 October 2024.
163 The Australian Financial Review, 'We can do it for under $100m': Start-up joins race to build local ChatGPT', 8 September 2025.
The Australian Financial Review, 'Meet Matilda, Australia's answer to ChatGPT', 14 August 2025.
Maincode, 'Maincode', accessed December 2025.
164 Ibid.

India has several Indic language initiatives, and Korea has both Naver and emerging competitors. It is relatively easy to adapt open-source models for local languages, enabling new entrants to compete. Furthermore, deployers can design systems where local models handle language input and output while reasoning tasks are routed to whichever provider offers the best performance, meaning local language capability captures only one layer of the value chain. These dynamics suggest that local language requirements may expand rather than narrow the competitive landscape.

It also remains to be seen whether this local language advantage will persist as a source of differentiation, or whether international FM providers will eventually match local performance with increasingly capable multilingual FMs.

2.4.1.5 Multi-homing

Multi-homing is a key feature of GenAI adoption in the APAC region. According to a Boston Consulting Group survey, around 20% of APAC firms describe their AI development strategy as using multiple FMs.[165]

Startups in the ASEAN region also deploy a range of models simultaneously. The survey finds that 76% use OpenAI models, 36% Llama, 33% models accessed through Hugging Face's open-source platform, 31% Claude, 15% Gemini, 14% Stability AI, 5% in-house models, 4% Cohere and 2% Mistral.[166] This implies an average of around two FMs per startup, though in reality this distribution could vary significantly.

The Indian Competition Commission also noted this multi-homing deployment pattern in its GenAI market study, finding that 43% of Indian GenAI startups prefer a hybrid model pipeline using both open-source and proprietary models.[167]

"Organisations demonstrate varied implementation strategies, with many enterprises opting for third-party solutions while simultaneously building internal capabilities." [168]

This multi-homing can have important implications for switching costs and lock-in, which we discuss in further detail in section 4.1.1.

2.4.1.6 Increasing use of small language models (SLMs)

Small language models (**SLMs**) with fewer parameters, streamlined designs and lower operating costs, are gaining traction, particularly for deployment on edge devices or cost-sensitive uses. Common examples include Microsoft Phi-4, GPT-4o mini, Gemma3 and smaller versions of Llama, Qwen 2.5 and DeepSeek-R1 models.[169] For deployers, these models offer a lower-cost alternative for tasks that do not require frontier model capabilities.

SLMs can be trained from scratch but can also be derived from larger models through pruning, distillation, quantisation or low rank techniques.[170] While SLMs have less general capabilities than leading LLMs, they can perform well where the model is adapted to domain-specific information. Wang et al. find that

---

165 BCG, 'In the Race to Adopt AI, Asia-Pacific Is the Region to Watch', 11 March 2025, p 7. This may underestimate the true scale of multi-homing as a further 59% of respondents replied that they mostly worked with large US tech firms, which may include some firms that work with multiple US FMs or utilise non-US models for a small percentage of their GenAI workloads.

166 GenAI Fund, 'ASEAN GenAI Startup Report 2024', 15 September 2024, p 23.

167 Competition Commission of India, 'Market Study on Artificial Intelligence and Competition', September 2025, p 23.

168 Ibid, p 46.

169 Hugging Face, 'Small Language Models (SLM): A Comprehensive Overview', 22 February 2025.
OpenAI, 'GPT-4o mini: advancing cost-efficient intelligence', 18 July 2024.

170 Pruning involves identifying low impact parameters and removing these from the model. Distillation involves teaching a smaller model to mimic a larger model's logic. Quantisation involves reducing the precision of mathematical representations by using lower precision data formats; and low-rank techniques represent high dimension matrices as the product of two lower dimension matrices, reducing parameter count.

fine-tuned SLMs perform well on tasks including question answering, coding solutions, recommendations and search ranking.[171]

The smaller size of these models reduces response times compared to LLMs. SLMs can also be deployed directly on devices such as smartphones or tablets rather than relying on cloud servers, which avoids network dependency and keeps sensitive information local. For example, Japan Airlines hosted a fine-tuned version of Microsoft's Phi-4 SLM on tablets to streamline cabin crew report filing in offline environments.[172]

These characteristics make SLM particularly relevant to APAC's mobile-first markets, where applications must often operate reliably on smartphones with limited or intermittent connectivity. Lower training and deployment costs for SLMs may also reduce barriers to entry for smaller, resource-constrained firms, potentially broadening the competitive landscape (see section 4.5.3).

2.4.2 Model deployment platforms

Market intermediaries such as open-source model repositories, cloud AI platforms, and service integrators play an important role supporting the distribution of FMs to application developers. These platforms and GenAI specialists also provide supporting tools to help adopters manage, adapt and integrate GenAI into their work. Global intermediaries shape GenAI competition through their APAC relationships with deployers and developers. In addition, multiple APAC-based firms such as Tencent, Alibaba and Baidu are active at this level of the value chain.

Given the varying levels of support provided to GenAI deployers, we classify these intermediaries as follows:

- Open-source model repositories and development hubs, such as Hugging Face and GitHub, allow developers to access a wide-variety of open-source models which can be adapted and deployed by internal teams. While these sites offer supporting tools, documentation and community support, they provide less dedicated support and therefore deployment success hinges on the developer's internal expertise.[173]
- Cloud AI platforms, such as Google Vertex AI, Amazon Bedrock, Tencent's TI Platform, Baidu's Qianfan and Samsung SDS FabriX provide unified interfaces for accessing and deploying models. These platforms typically offer access to multiple FMs through unified APIs, along with supporting tools for machine learning operations, infrastructure management and orchestration.[174] This can simplify switching between models.
- Service integrators, such as TCS, offer external expertise and solutions for GenAI adopters with limited in-house technical skills. These firms can supply and install pre-packaged GenAI solutions or work with internal stakeholders to design customised solutions. Integrators can also take on an even wider role, managing supplier relationships and procurement, or overseeing broad deployment strategies and customer migrations between providers.

A brief overview of key intermediaries and their core functionalities is outlined in Figure 7 below.

---

171 Wang et al., 'A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness', 4 November 2024.
172 Fujitsu, 'Fujitsu and Headwaters trial on-device generative AI solution to streamline JAL cabin crew workflows', 10 April 2025.
173 Hugging Face, 'Hugging Face', accessed November 2025.
174 Google Cloud, 'Innovate faster with enterprise-ready AI, enhanced by Gemini models', accessed November 2025.
CloudOptimo, 'Amazon Bedrock vs Azure OpenAI vs Google Vertex AI: An In-Depth Analysis', 5 May 2025.
AWS, 'Amazon Bedrock', accessed November 2025.

Figure 7: Key features of prominent GenAI deployment platforms active in APAC

Google Vertex AI

- Access to Google, third-party proprietary, and open-source models thorough **Model Garden**.
- **Vertex AI Pipelines** for workflow orchestration.
- **Vertex AI Training** and **Prediction** features to optimize model training and identify efficient infrastructure set ups.
- **Vertex AI Evaluation** tools to identify best model for specific workloads.

Amazon Bedrock

- **Unified API** access to over 100 foundation models.
- **Prompt caching and prompt routing** features to reduce running costs.
- **Bedrock Guardrails** to block harmful content and minimize hallucinations.
- **Knowledge Bases** for RAG workflows **Bedrock Data Automation** tools to build data pipelines that utilize proprietary data for fine-tuning.

Baidu Qianfan

- Provides **unified access** to Baidu's ERNIE FMs, along with APIs and tools for **fine-tuning**.
- Includes a **multi-agent framework** that supports agent orchestration, parallel task execution, progress tracking, and pre-configured agents for web search, coding, and report generation.
- **End-to-end application development** through retrieval-augmented generation, agent configuration, workflow management, and ongoing monitoring.

Alibaba Cloud Model Studio

- Offers **industry-leading** foundation models including Alibaba's Qwen models and third-party models.
- **Full-chain development** to accelerate application development.
- **Enterprise-grade security** to reduce privacy risks in deploying applications.

Source: RBB summary of key platform features based on:
Google, 'Vertex AI Platform', accessed November 2025.
AWS, 'Amazon Bedrock', accessed November 2025.
Baidu, 'Qianfan Large Model Platform', accessed November 2025.
Alibaba Cloud 'Alibaba Cloud Model Studio', accessed November 2025.

Though these platforms and integrators differ in their target users, they share several characteristics relevant to the competitive dynamics of the GenAI market.

- **Access to multiple competing models:** Cloud AI platforms typically provide access to a wide range of models, providing customer choice. This is despite the fact that many platform operators are vertically integrated and could restrict access to third-party or open-source models to favour their own models. For example, Google Vertex AI offers over 200 FMs including Anthropic's Claude and Meta's Llama alongside Google Gemini and Amazon Bedrock provides standardised API access to hundreds of FMs.[175] APAC players display similar behaviour: Alibaba's Cloud Model Studio offers access to Qwen, DeepSeek and Kimi models, while Tencent's TI platform supports models from Meta, Mistral AI, Baichuan AI and Zhipu AI alongside Tencent's own Hunyuan models.[176] Service integrators similarly form distribution partnerships with multiple FM suppliers to cater to their customer preferences. Through its partnerships with AWS Bedrock, Google Cloud Vertex AI and Microsoft Azure, TCS provides access to models from OpenAI, Anthropic, Meta, Alibaba and DeepSeek.[177]
- **Model evaluation tools:** These platforms provide model evaluation tools as a core functionality to help deployers assess model performance against their requirements and identify the optimal FM for their

[175] Ibid.
[176] Tencent Cloud, 'What mainstream large language models does the large model application building platform support?', 31 July 2025.
[177] TCS, 'TCS Partners with Google Cloud to Integrate Gemini Enterprise for its Workforce and Customers', 14 October 2025.
TCS, 'TCS to Build New AI-Led Solutions for Business Transformation in Collaboration with Microsoft', 20 June 2025.
TCS, TCS launches new Generative AI practice in collaboration with AWS', 27 November 2023.
For the list of FMs available in AWS Bedrock, see 'AWS'.
For the list of FMs available in Microsoft Azure, see 'Azure'.
For the list of FMs available in Google Cloud Vertex AI, see 'Google Cloud'.

use case. For instance, Amazon Bedrock offers multiple evaluation programmes, including the use of LLM judges or human reviewers, against benchmarks such as correctness, completeness and harmfulness of responses.[178]

- **Customisation and fine-tuning tools:** Cloud AI platforms can incorporate internal company data to tailor models for specific use cases, such as training a model on internal documents so it can answer company-specific questions. Platforms provide tools and infrastructure to manage this training and adjust model settings.

Recent feature releases on these platforms have focused on incorporating GenAI advances relevant to enterprise use cases. Retrieval-Augmented Generation (**RAG**) tools, now available in Amazon Bedrock, Google Vertex AI Search and Alibaba Cloud, allow models to pull in relevant company data when generating responses, improving accuracy and reducing hallucinations.[179] Platforms are also releasing agent development tools that enable multi-step automated workflows: for instance, an agent that can research a query, draft a response, and send it for approval, with orchestration tools coordinating multiple agents working on different tasks.[180]

Platform operators are also expanding their model offerings. Since October 2025, Google Vertex AI has added new models from seven different third-party FM developers to its model garden, including Alibaba, DeepSeek, MiniMax, and Kimi, broadening access to APAC-developed models for deployers using the platform.[181] Google Vertex AI and Amazon Bedrock continue to expand their APAC presence. In particular, Amazon Bedrock recently launched in Thailand, Malaysia and Taiwan.[182]

The availability of these platforms in APAC matters for competitive dynamics. By providing access to multiple models through unified interfaces, they make it easier for deployers to evaluate alternatives and switch between providers. This is discussed further in section 4.4.

2.4.3 Applications

2.4.3.1 Introduction

APAC consumers and businesses are rapid adopters of GenAI. According to a Deloitte survey, almost 70% of APAC consumers use GenAI, with markedly higher usage rates in developing economies such as India and Southeast Asia (over 70%), compared to Australia, New Zealand and Japan (40-55%).[183] Enterprises adoption is also strong: 55% of APAC companies have implemented GenAI initiatives, compared to 46% of their North American and 42% of their European counterparts.[184]

FM development requires significant capital expenditure, meaning most GenAI startups focus on developing applications rather than building foundation models. In ASEAN, 91% of GenAI startups report developing GenAI applications compared to 20% building GenAI platforms and 19% focusing on FMs.[185] Indian startups show a similar pattern: 67% develop GenAI apps while only 3% focus on FM development.[186]

---

178 AWS, 'Amazon Bedrock Evaluations', accessed November 2025.
179 AWS, 'Retrieve data and generate AI responses with Amazon Bedrock Knowledge Bases', accessed November 2025.
Google Cloud, 'Vertex AI RAG Engine overview', accessed December 2025.
Alibaba Cloud, 'Dash Vector', accessed November 2025.
180 AWS, 'Amazon Bedrock Agents', accessed November 2025.
Tencent Cloud, 'What is the Agent Development Platform', 16 September 2025.
181 Google Cloud, 'Vertex AI release notes', accessed December 2025
182 AWS, 'Amazon Bedrock now available in the Asia Pacific (Thailand, Malaysia, and Taipei) Regions', 29 September 2025.
183 Deloitte, 'Generative AI in Asia Pacific: Young employees lead as employers catch up', 13 May 2024, Figure 3.
184 Infosys, 'Generative AI Radar APAC', 2024, p 5.
185 GenAI Fund, 'ASEAN GenAI Startup Report 2024', 15 September 2024, p 16.
186 Competition Commission of India, 'Market Study on Artificial Intelligence and Competition', September 2025, pp 12-13.

2.4.3.2 GenAI in super apps

Super apps (platforms that integrate multiple services such as online payments, messaging, ride-hailing, and shopping, into a single application) are more prevalent in APAC than in Western markets and create distinct distribution dynamics for GenAI. Because these platforms span many different functions, they offer multiple integration points for AI features: search, customer service, recommendations, content curation and merchant tools can all be enhanced within one ecosystem. This breadth generates diverse user data from transactions, messaging and browsing that can inform AI personalisation.

GenAI integrations typically enhance existing platform functions. WeChat's AI Search uses GenAI to interpret queries and return relevant results from across its ecosystem, including articles, videos and mini apps (services which run inside the ecosystem).[187] Tencent, owner of WeChat/Weixin, recently tested integrating DeepSeek-R1 alongside its own Hunyuan model to improve WeChat/Weixin search quality.[188] LINE has embedded a GenAI chatbot directly into its messaging interface, while Paytm in India has partnered with Perplexity to offer AI-powered financial search.[189] Alibaba deploys GenAI on both sides of its Taobao e-commerce platform (which can be accessed as a mini app from the Alipay super app): the Wenwen chatbot recommends products to consumers, Taobao's Business Advisor function offers merchants insights into industry trends and customer demand, and the Ali Xiaomi customer service chatbot answers questions about deliveries and returns.[190] The use of GenAI even extends into the healthcare domain as the Ant Group's AI healthcare manager, also available as an Alipay mini app, offers over 100 AI-powered services, including doctor recommendations, medical report analysis and personalised health advice.[191]

The mobile-first nature of these super apps has driven investment in voice-based and image-based AI.

- Gojek's assistant Dira helps customers navigate to various banking functions in response to voice commands, while Gojek's rival Grab incorporates voice commands in the AI Driver companion tool that allows delivery drivers to safely flag road conditions to the wider driver network.[192] This technology is also used as a consumer-facing voice assistant that allows visually impaired users to book Grab rides with voice prompts.[193]
- Alipay's image search allows users to scan pictures for translation or to retrieve information, while Taobao offers AI photo editing tools for merchant marketing content.[194]

Super app ecosystems also provide distribution channels for third-party GenAI applications through mini apps. Luckin Coffee's Alipay mini app uses an AI chatbot with voice recognition that simplifies the ordering process, connects to Alipay's mobile payment system, and enables browsing, ordering, and payment through a single user interface.[195]

The integration of GenAI in super apps has important implications for product distribution for external FM developers and GenAI application providers. We discuss this further in section 4.1.3.

---

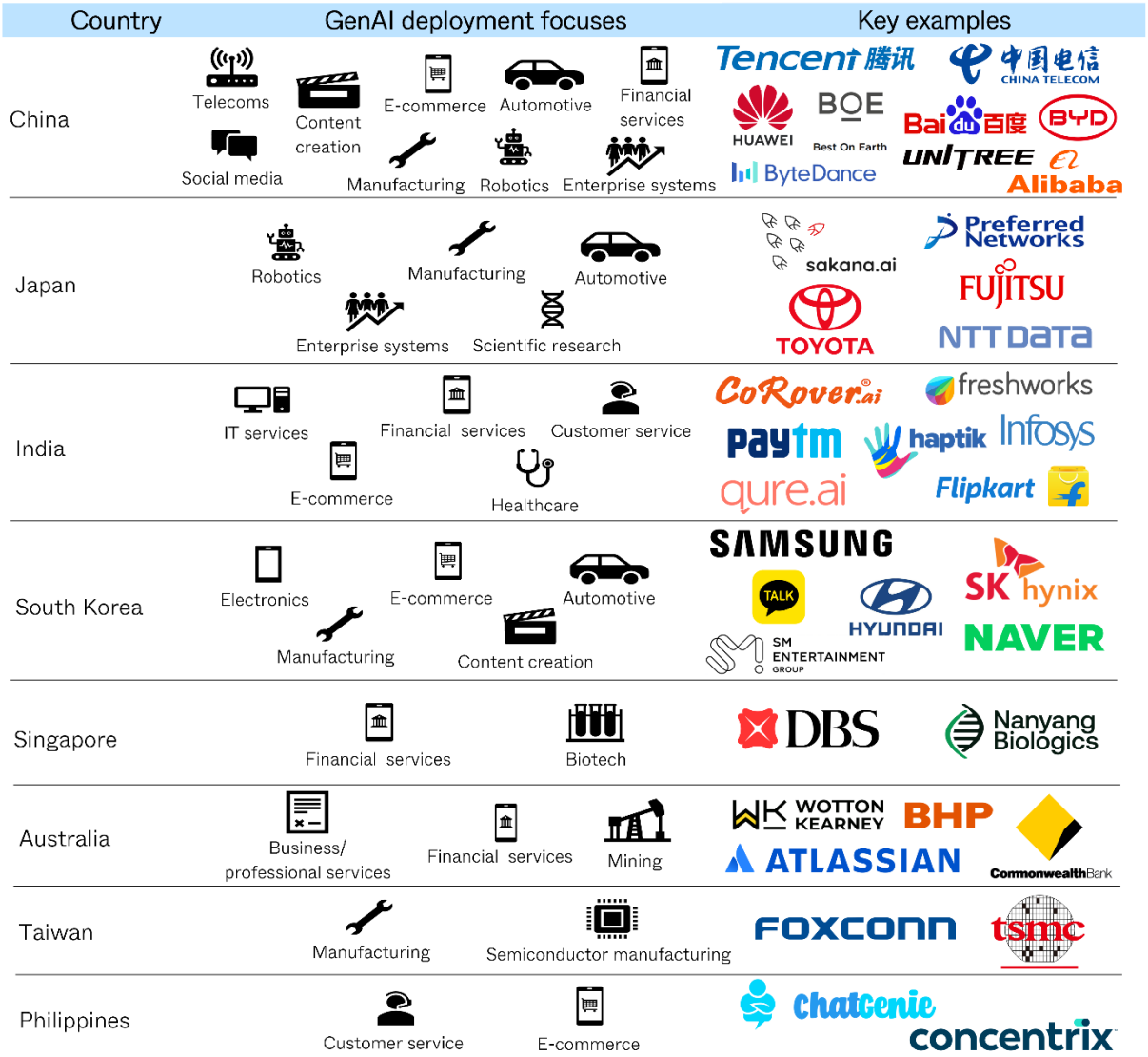
187 ITC, 'WeChat and WeCom Integrating DeepSeek: What Brands Need to Know', accessed November 2025.
188 ITC, 'WeChat and WeCom Integrating DeepSeek: What Brands Need to Know', accessed November 2025.
Reuters, 'Tencent's Weixin app, Baidu launch DeepSeek search testing', 17 February 2025.
189 Line, 'LINE AI assistant brings AI power to LINE messenger', 22 February 2024.
Paytm, 'Paytm Partners with Perplexity to Bring the Power of AI to Crores of Indians', 27 February 2025.
190 Alizila, 'Taobao and Tmall Upgrades Consumer Shopping Experience and Merchant Support Through AI', 13 June 2024.
191 Business Wire, 'Ant Group's AI Healthcare Manager AQ Hits 100 Million Users', 28 July 2025.
192 PR newswire, 'GoTo Group and Google Cloud Extend Collaboration on Generative AI with Groundbreaking In-app Voice Assistant, Dira', 24 September 2024
Grab, 'Grab deploys agentic AI to empower merchants and driver partners', 8 April 2025.
193 Vulcan Post, 'Grab launches its first AI Centre of Excellence, plans to hire at least 50 "high-value" roles', 23 May 2025.
194 Yahoo Finance, 'Alipay adds AI image search to enhance 'super app' role in rivalry with Tencent's WeChat', 31 December 2024.
Alizila, 'Taobao and Tmall Upgrades Consumer Shopping Experience and Merchant Support Through AI', 13 June 2024.
195 Tech in Asia, 'Alipay, Luckin Coffee launch conversational AI payments in China', 11 September 2025.

2.4.3.3 Application focus areas

Beyond super apps, GenAI application development in APAC tends to reflect existing industrial strengths. Developers focus on sectors where domain expertise is available and use cases are well understood.

Figure 8 below illustrates this variation across APAC countries. China's scale and closed ecosystem have supported specialised industrial applications in manufacturing and robotics, and general-purpose chatbots built on domestic FMs.[196] Japanese developers concentrate on robotics, automotive and manufacturing, building on the country's established industrial base. Indian developers target IT services, customer service and financial services, where they have outsourcing strength and domestic capacity. Developers in South Korea focus on electronics, content creation and automotive. Singapore's applications cluster around financial services and biotech, while Taiwan's focus on semiconductor and electronics manufacturing reflects its position in global supply chains.

Figure 8: Summary of key GenAI deployment use cases



Source: RBB analysis based on desk research of publicly available sources.

[196] Wired, 'Meet the Chinese Startup Using AI – and a Team of Human Workers – to Train Robots', 5 November 2025.
NVIDIA, 'Following the Prompts: Generative AI Powers Smarter Robots With NVIDIA Isaac Platform', 8 January 2024.

2.4.3.4 Standardisation and interoperability

Several standardisation and interoperability initiatives have emerged globally and are being adopted by developers across APAC. Firms such as Google and Anthropic have released open-source standards addressing different interoperability challenges.

Anthropic's Model Context Protocol (**MCP**) connects FMs to external data sources, providing a unified API and translation features to convert different data formats into a common format for parsing by FMs.[197] Following its November 2024 introduction and subsequent endorsement by OpenAI and Microsoft, MCP was adopted by major Chinese platforms.[198] However, MCP adoption currently remains geographically uneven. While community-developed MCP servers exist for Naver APIs (created by third-party developers), official platform adoption has not occurred. Google's Agent-to-Agent standard (**A2A**) addresses a different challenge, offering an open communication standard to enable AI agents from different developers and platforms to interact and collaborate with each other.[199]

Orchestration tools take a different approach: rather than getting providers to adopt common protocols, they provide a translation layer that lets developers interact with different FM providers in the same way:

- **LangChain** provides a development framework that lets developers write code once and deploy it across different FM providers. The framework has achieved over 130 million downloads globally and supports over 70 model providers.[200]
- **OpenRouter** offers API access to multiple FMs through a single interface, allowing developers to switch between providers without rewriting their code. Its smart routing feature can automatically select the best FM for a given query based on real-time performance data.[201]
- **Hugging Face** provides a single API framework allowing access to any open-source model hosted on its platform.[202]

Some FM providers have also adopted OpenAI-compatible APIs, allowing developers to switch from OpenAI with minimal code changes. For example, Moonshot AI's API mirrors OpenAI's format, making switching between the two FMs seamless.[203]

The adoption of these tools suggests developers value flexibility when choosing which FMs to use. As discussed further in section 4.4.2, these tools facilitate FM evaluation and switching, fostering competition between FM providers.

---

197 IBM, 'What is Model Context Protocol (MCP)?', accessed November 2025.
Google Cloud, 'What is the MCP and how does it work?', accessed December 2025.
198 Techzine, 'How the Model Context Protocol has taken the AI world by storm', 26 April 2025.
SCMP, 'Chinese tech giants race to expand AI services market with latest open-standard protocol', 21 April 2025.
AIBase, 'Alibaba Announces Full Support for MCP Protocol', 9 April 2025.
199 Google, 'Announcing the Agent2Agent Protocol', 9 April 2025.
200 LangChain, 'LangChain's Second Birthday', 24 October 2024.
201 OpenRouter, 'The Unified Interface for LLMs', accessed November 2025.
OpenRouter, 'Provider Routing', accessed December 2025.
202 Hugging Face, 'Inference Providers', accessed November 2025.
203 Moonshot AI, 'Migrating from OpenAI to Kimi API', accessed November 2025.
China Daily, 'US firm's AI tech ban set to inspire homegrown innovation', 25 July 2024.

3 Possible GenAI competition concerns

3.1 Relevant considerations in assessing GenAI competition

This section identifies potentially relevant competition concerns for GenAI.[204] Given the rapid adoption of GenAI, competition authorities in APAC have published reports and market studies examining competition in the industry. Regulators in South Korea, Japan, Taiwan and Australia have investigated the GenAI market specifically, while regulators in New Zealand, India and Singapore have examined the wider AI market.[205] Regulators in China are also examining competition issues arising from other AI and algorithm-driven markets, as evidenced by the State Administration for Market Regulation's (**SAMR**) recent draft antitrust compliance guideline for internet platforms.[206] Meanwhile, regulators in other regions, including the UK, EU and US have examined possible competition concerns that might arise from GenAI.[207]

We draw on these reports to identify concerns raised by regulators, grouping them into five categories. For each category, we discuss potential efficiencies and pro-competitive effects alongside the concerns. We conclude with a summary of concerns identified by APAC regulators specifically. Section 4 then examines whether these concerns have materialised to date in APAC.

3.1.1 Vertical integration

A vertically integrated firm operates at multiple levels of a supply chain, bringing different parts of the production process in-house rather than relying on external suppliers.[208] In GenAI, this is particularly relevant because the "stack" is complex, including foundation models, deployment platforms, and downstream applications. As a result, integration decisions by major GenAI providers can influence the competitive landscape for rivals who rely on these same layers.

Potential benefits of vertical integration

Vertical integration can drive efficiency and innovation by allowing a firm to coordinate decisions across the stack. Key benefits include:

- **Lower prices through the elimination of double marginalisation:** When firms operate separately at different levels the supply chain, both the upstream and downstream firms may add their own separate markups to the cost. This can lead to higher prices for the final consumer. A vertically integrated firm solves this issue by bringing both stages under common ownership: with only a single profit markup, the integrated firm can offer lower final prices.[209]

---

204 The analysis draws on the framework provided in RBB, 'Competitive Dynamics of Generative AI', 12 June 2025, section 3.

205 KFTC, 'Generative AI and Competition', 18 December 2024.
JFTC, 'Report regarding Generative AI ver. 1.0', 6 June 2025.
TFTC, 'Competition Law Issues Related to Generative Artificial Intelligence, consultation paper', 7 September 2025
ACCC, 'Digital platform services inquiry, final report', 13 March 2025.
CCI, 'Market study on artificial intelligence and competition', September 2025.
NZCC, 'Navigating the rise of AI: Perspectives from a competition and consumer regulator', 7 July 2025.
OECD, 'Artificial Intelligence, Data and Competition – Note by Singapore', 29 May 2024.

206 China Daily, 'SAMR unveils draft antitrust guideline for internet firms', 18 November 2025.
SCMP, 'China targets AI-powered price manipulation in new antitrust guidelines', 17 November 2025.

207 RBB, 'Competitive Dynamics of Generative AI', 12 June 2025, section 3.
Competition and Markets Authority, 'AI FMs: Initial Report', 18 September 2023.
Competition and Markets Authority, 'AI FMs: Technical Update Report', 16 April 2024.
European Commission, 'Competition Policy Brief. Competition in Generative AI and Virtual Worlds', 18 September 2024.
Autorité de la concurrence, 'On the competitive functioning of the generative artificial intelligence sector', 28 June 2024.
Autoridade da Concorrência, 'Competition and Generative Artificial Intelligence', November 2023.
Federal Trade Commission, 'Partnerships Between Cloud Service Providers and AI Developers', January 2025.

208 Bishop, S. & Walker, M., 'The Economics of EC Competition Law: Concepts, Application and Measurement', 2003, p 189.

209 Spengler, J.J, 'Vertical Integration and Antitrust Policy', Journal of Political Economy, Vol. 58, No. 4, August 1950, pp 347-352.

- **Better product integration and performance:** An integrated firm can optimise products to work seamlessly together.[210] For example, a firm developing both FMs and downstream applications can engineer the model specifically to reduce latency for that application.
- **Faster, coordinated innovation:** GenAI development and deployment require significant upfront investment**.** When firms are separate, an upstream model developer might hesitate to customise their technology for a specific downstream application, fearing the partner might later renegotiate terms or fail to launch the product successfully. Vertical integration solves this coordination problem.[211]

However, full integration is not always necessary. Firms may instead seek to achieve similar benefits through vertical agreements, such as long-term partnerships or exclusive licensing deals (see section 3.1.4).

Potential competition concerns

While generally efficient, vertical integration can harm competition if a firm with significant market power uses its position in one layer to harm competition in another. Theories of harm potentially relevant in a GenAI context include:

- **Input foreclosure:** A vertically integrated firm which operates at different levels of the value chain could restrict or degrade access to upstream inputs for downstream competitors.[212] For example, a firm that develops both FMs and applications could worsen API access terms or increase prices for rival GenAI application developers, weakening their ability to compete with the integrated firm's own applications.
- **Customer foreclosure:** A vertically integrated firm may restrict access to important sales channels, preventing upstream rivals from reaching end users.[213] For example, a super app with hundreds of millions of users could refuse to integrate third-party FMs, or a deployment platform could limit FM distribution to its own proprietary models.
- **Self-preferencing:** A vertically integrated firm may systematically favour its own products over rivals' products.[214] In GenAI, this might involve pre-installing proprietary GenAI solutions on devices or restricting user data to internal teams rather than sharing it with rival FM developers.[215] Self-preferencing often overlaps with input and customer foreclosure, and captures a range of conduct that may arise in digital markets.

The assessment of whether vertical integration creates net benefits or competitive harms requires careful case-by-case analysis, weighing efficiency gains against potential foreclosure and self-preferencing concerns.

3.1.2 Platform and ecosystem dynamics

Many firms operating in the GenAI value chain are active across multiple **adjacent** markets. For example, a cloud provider or a productivity software firm may also offer GenAI applications. Competition authorities have raised concerns about how incumbent platforms might influence the market through bundling, ecosystem effects and first-mover advantages.[216] However, these ecosystems can generate economic efficiencies.

---

210 Riordan, M.H. & Salop, S.C., 'Evaluating Vertical Mergers: A Post-Chicago Approach', Antitrust Law Journal, Vol. 63, No. 2, 1995, pp 513-568.
211 Liu, X., 'Vertical integration and innovation', International Journal of Industrial Organization, Vol. 47, July 2016, pp 88-120.
212 ACCC, 'Digital platform services inquiry, final report', 13 March 2025, p 14.
213 Hart, O. & Tirole, J., 'Vertical Integration and Market Foreclosure', Brookings Papers on Economic Activity: Microeconomics, Vol. 1990, 1990, pp 205-286.
214 Padilla, J., Perkins, J. & Piccolo, S., 'Self-Preferencing in Markets with Vertically-Integrated Gatekeeper Platforms', Journal of Industrial Economics, Vol. 70, No. 2, June 2022, pp 371-395.
215 KFTC, 'Generative AI and Competition', 18 December 2024, p 72.
216 KFTC, 'Generative AI and Competition', 18 December 2024, section 3.2.1.2 and section 3.1.3.

Potential benefits from integration of GenAI with platforms

When a firm integrates GenAI tools into its existing platforms, it can create value that standalone firms cannot easily match:

- **Lower barriers to entry:** Incumbents with a presence in one market can use that existing relationship to successfully launch products in a new, adjacent market. For example, a firm with a popular productivity suite can reach millions of daily users. This allows the firm to bypass the costly, slow process of acquiring customers from scratch, which can be a hurdle for standalone providers that must build brand awareness.
- **Better user experience:** The value of GenAI is often maximised when it works in tandem with other tools. A firm that operates across adjacent markets can engineer its products to interoperate seamlessly, such as an AI assistant that drafts emails directly within an email application or sets up meetings and adds them to a calendar. This is significantly more efficient than the alternative arrangement, where a user must manually copy and paste text between a "chatbot" website and their separate email application.
- **Economies of scope:** It can be efficient for a firm to use the same inputs across multiple distinct products.[217] For example, an incumbent can build one FM and then fine-tune it for different applications (like a coding assistant or a chatbot) at a fraction of the original cost. Additionally, they can run all these different tools on the same shared cloud infrastructure (servers and chips) to maximise utilisation.

Potential competition concerns

- **Tying and bundling in adjacent markets:** Conglomerate firms operate across multiple related markets, and there is a risk that an incumbent may bundle its new GenAI products with existing core services in ways that weaken competitors' ability to compete or create barriers to entry.[218] This can occur through pure bundling (where products are only sold together), mixed bundling (where the bundle is cheaper than buying items separately), or tying arrangements where a product can only be purchased if the customer also purchases another product. By linking a nascent GenAI tool to a "must-have" product, such as a widely used operating system or productivity suite, an incumbent can capture a large share of the market. For example, if an incumbent includes its own GenAI writing assistant in a standard office software subscription, corporate customers may have little incentive to pay extra for a competing tool from a standalone rival. This practice can reduce the addressable market for standalone rivals, who may struggle to compete even with superior technology because they cannot match the reach or pricing of the incumbent's bundle. However, this strategy is only likely to harm competition if: (i) the incumbent has significant market power in the "must-have" product; (ii) the bundle denies rivals the scale they need to compete effectively; and (iii) it thereby forecloses rivals who offer a product of comparable quality and price.
- **Ecosystem entrenchment and competitive constraints:** Firms offering a suite of connected products can create "walled gardens" that are difficult for users to leave.[219] This strategy often reflects good product design: consumers value the convenience of seamless data sharing and interoperability between their apps and devices. However, this integration can increase switching costs: a user might prefer a rival's model for one specific task, but if switching means losing their data, they may be more likely to stay within the incumbent's ecosystem. This can lead to a competition concern if the ecosystem

---

217 Bishop, S. & Walker, M., 'The Economics of EC Competition Law: Concepts, Application and Measurement', 2003, p 467 and p 916.

218 Adams, W.J. & Yellen, J.L., 'Commodity Bundling and the Burden of Monopoly', Quarterly Journal of Economics, Vol. 90, No. 3, 1 August 1976, pp 475-498.
McAfee, R.P., McMillan, J. & Whinston, M.D., 'Multiproduct Monopoly, Commodity Bundling, and Correlation of Values', Quarterly Journal of Economics, Vol. 104, No. 2, 1 May 1989, pp 371-383.
ACCC, 'Digital platform services inquiry, final report', 13 March 2025, p 316.

219 TFTC, 'Competition Law Issues Related to Generative Artificial Intelligence, consultation paper', 7 September 2025, p 7.

provider holds significant market power and a rival with a superior individual product cannot compete because they cannot replicate the incumbent's entire ecosystem.

- **First-mover advantages and market tipping**: In digital markets, early innovators can entrench their position, for instance by rapidly accumulating user data and securing default positions on devices. Once customers become familiar with a specific interface, they may be reluctant to switch, making it difficult for later entrants to gain a foothold.[220] In a functioning market, this is simply the reward for innovation: the first firm to solve a consumer problem wins their business. However, this dynamic can lead to a competition concern if equally efficient rivals cannot catch up, and the market "tips" to the incumbent.

Assessing the competitive effects of platform integration requires balancing clear efficiency gains (like seamless user experiences and lower entry barriers) against the risk of harm to competition. Crucially, if the risk of "tipping" discussed above materialise, the market structure may become concentrated as discussed in the next section.

3.1.3 Concentration concerns

The potential for platform network effects and "feedback loops" discussed above raises a broader structural question for GenAI. While the current landscape is fragmented, with numerous suppliers at every level of the supply chain, regulators in APAC and globally are assessing whether this is merely a transitory phase. Their concern is that, should these self-reinforcing effects materialise in the GenAI sector, the market could eventually consolidate around a small number of players.[221] In particular the UK's Competition and Markets Authority refers to potential advantages gained from increasing returns to data scale, high levels of personalisation, and platform-level network effects as potential drivers.[222]

Potential benefits of consolidation and scale

Scale can drive significant consumer benefits:

- **Competition on quality:** Scale can be a powerful driver of product improvement. If data feedback loops prove relevant in GenAI, a supplier with more users can gather more data to fine-tune its models and personalise results, creating a better experience than a smaller rival could offer. Likewise, **should** platform-level network effects materialise, larger platforms could offer greater utility simply by connecting more participants across distinct groups. For example, a large user base attracts more third-party developers, giving users a wider choice of compatible tools (cross-side network effects). Simultaneously, a larger community can directly improve the platform through better peer support and collaboration features (same-side network effects), making the service more valuable for everyone involved.

- **Combination of technical expertise:** Consolidation allows firms to pool scarce technical expertise and R&D budgets. Knowledge sharing can accelerate innovation, as combined teams can often solve complex problems faster than fragmented ones. It also reduces the waste of duplicative investments in R&D, where multiple firms spend resources solving the same challenges. These efficiencies can be passed on to consumers in the form of lower prices or better products.

- **Improved bargaining position:** Consolidation can strengthen a firm's ability to negotiate with powerful suppliers or customers. A larger, combined entity may obtain better trading terms. These savings may be passed on in lower prices for consumers.

---

220 KFTC, 'Generative AI and Competition', 18 December 2024, section 3.1.3.
221 NZCC, 'Navigating the rise of AI: Perspectives from a competition and consumer regulator', 7 July 2025, pp 4-5.
CMA, 'AI Foundation Models Update paper ', 11 April 2024, pp 20-21.
222 CMA, 'AI Foundation Models Update paper ', 11 April 2024, pp 20-21.

- **Market standardisation and interoperability:** Consolidation can lead to standardised technical systems. This reduces fragmentation, making it cheaper and easier for third-party developers to build compatible tools.

Potential competition concerns

- **Market tipping:** As discussed above, tipping occurs when a market consolidates to a single provider or a small group of large players because their advantage becomes self-reinforcing.[223] For example, if a model improves significantly with every additional user interaction, the leading firm may eventually pull so far ahead that rivals cannot catch up. This dynamic can be exacerbated by high switching costs or platform lock-in, making it difficult for users to leave. However, this dynamic is only likely to harm competition if the leading firm's advantage cannot be replicated. If models are trained on widely available public data (as is currently common), rivals can replicate the leader's success, meaning high concentration may be temporary.[224]

Given the significant number of suppliers currently active in GenAI, regulators must be careful not to mistake a high market share for market failure. Absent significant consumer lock-in or high barriers to entry and expansion that cement the leader's position, the rapid growth of a leading firm is consistent with healthy competition in innovation. As such, any assessment of tipping risks must focus on whether strong platform-level network or data feedback effects are likely to occur, and whether the underlying inputs remain accessible enough for competing firms to challenge the market leader.

3.1.4 Agreements between firms

The GenAI market is characterised by numerous commercial partnerships and collaborations between firms across the value chain. While these agreements typically generate efficiencies and foster innovation, they can occasionally lead to anticompetitive effects. Consequently, APAC competition authorities are monitoring these partnerships to ensure they do not restrict competition.

Potential benefits of agreements between firms

- **Innovation:** Agreements allow firms to exchange capital, technical know-how and technology. This collaboration can accelerate innovation by combining strengths.[225] Formal contracts can also solve the "hold-up" problem: without an agreement, a firm may be reluctant to invest in building a product dependant on a partner's technology, fearing the partner could later change access terms. Formal agreements can provide the certainty needed to engage in such investment.

- **Lower barriers to entry:** Developing GenAI products can require large upfront capital investment. Partnerships can allow firms to overcome resource and technical hurdles.[226] For example, a specialised software supplier could partner with a model provider to launch product without having its own research team, while the model provider gains access to a new market.

- **Access to large user bases and accelerated time to market:** Partnerships can provide smaller players with access to scale. For example, a specialised virtual assistant developer can integrate its tool into a widely used productivity suite or a popular operating system and access millions of users.

---

223 Bishop, S. & Walker, M., 'The Economics of EC Competition Law: Concepts, Application and Measurement', 2003, p 416.

224 Hagiu, A. & Wright, J., 'Artificial intelligence and competition policy', January 2025, pp 5-9. We note that preliminary analysis by Hagiu and Wright (2025) finds that concerns of strong cross-user data feedback loops are unlikely to materialise at the FM level given that models are mainly trained on widely available or reproducible data, with limited usage of user data for training purposes. Additionally, platform-level network effects may be limited as AI applications are not distributed via closed app stores which only promote apps built on proprietary FMs. The authors find that there may be within-user data feedback loops which could allow personalised services within GenAI applications. However, this would be more likely to provoke user lock-in and entrench the advantages of larger players without necessarily driving this consolidation towards larger players in the first place.

225 ACCC, 'Digital platform services inquiry, final report', 13 March 2025, p 304.

226 JFTC, 'Report regarding Generative AI ver. 1.0', 6 June 2025, pp 36-37.

- **Competitive pressure in adjacent markets:** Agreements between firms can introduce new competition into established sectors. When a firm integrates a novel GenAI solution (such as an advanced customer service chatbot) into its existing offering, it improves the customer experience. This forces rivals in that market to respond with their own innovations or price cuts, effectively raising the competitive standard for the whole industry.

Potential competition concerns

- **De facto control and information exchange:** Specific structures can raise concerns if they allow a firm to influence a rival's strategy without triggering a merger review. This merger effect occurs if a minority investment or agreement grants the investor de facto control, effectively stopping the partner from competing aggressively. Additionally, partnerships can require sharing sensitive data like product roadmaps and could provide access to sensitive information on rivals' commercial strategies or technical developments through their partner's relationship with these competitors. This can be a competition concern if it grants the investor decisive influence over the partner, otherwise reduces incentives to compete, or increases the risk of collusion.
- **Exclusive dealing:** Firms may use exclusivity clauses to lock up key parts of the supply chain, preventing rivals from accessing essential inputs or customer bases.[227] For example, a GenAI integrator or a popular deployment platform might enter an agreement to only distribute one developer's FMs, effectively cutting off other developers from that route to market. This can raise competition concerns if the agreements cover a significant share of the market and exclusive deals deny rivals the scale they need to compete effectively.
- **Algorithmic collusion:** This concern is distinct from the agreements above because it does not necessarily involve agreement between firms, but rather an outcome where AI tools can cause market prices to align. As firms delegate pricing to GenAI agents, two risks emerge: "hub-and-spoke" coordination (where competitors use the same third-party algorithm) or autonomous learning (where agents independently learn that price wars are unprofitable).[228] This outcome is more likely to materialise in markets characterised by high transparency, homogeneous products, and frequent interactions, where algorithms can accurately monitor rivals and predict that aggressive competition will reduce long-term profits.

The assessment of whether partnerships create net benefits or harms requires careful analysis on a case-by-case basis. This involves considering the potential for increased innovation and expanded access to resources against possible competition risks.

3.1.5 Mergers

Merger control policy is well-equipped to assess combinations between firms in all industries, including GenAI. GenAI has raised some specific concerns related to "killer acquisitions" and "talent acquisitions".

Potential benefits of mergers

- **Accelerated commercialisation of new solutions through integration and scale**: Acquisitions can generate pro-competitive benefits by combining a startup's innovative ideas with an incumbent's resources. Small GenAI firms often face significant constraints in capital, computing power and distribution. A merger can lead their technology to be integrated into established products. This can deliver better features and greater choice to consumers faster than the startup could achieve alone.

---

227 CCI, 'Market study on artificial intelligence and competition', September 2025, p 98.
228 ACCC, 'Digital platform services inquiry, final report', 13 March 2025, p 325.
JFTC, 'Report regarding Generative AI ver. 1.0', 6 June 2025, p 38.

- **Efficient allocation of talent:** Acquisitions can promote efficiency by moving skilled teams to where they can be most productive. Absorbing a high-quality team into a firm with the resources to support them can allow technical expertise to be deployed more effectively.

Potential competition concerns

- **Killer and reverse killer acquisitions:** In a killer acquisition, the incumbent buys a promising rival to shut it down, preventing it from growing into a competitor. In a reverse killer acquisition, the incumbent buys a rival to discontinue its *own* internal innovation efforts.[229] In both cases, future competition may be lost and longer-term innovation incentives may be reduced, which could result in poorer quality and reduced consumer choice. These theories of harm apply only if the target firm or the incumbent was likely to grow into a significant competitive constraint.
- **"Acqui-hires":** Competition authorities are concerned about transactions structured to transfer a startup's core team to an incumbent without acquiring the corporate entity itself.[230] This practice may replicate the anticompetitive effects of a merger (that is, eliminating a competitor and increasing market power) while potentially falling outside standard merger review thresholds because no shares or "assets" were technically purchased. Some regulators are examining whether, and how, their merger control regimes could be applied to capture this kind of practice.[231]

While acquisitions in the GenAI sector are likely to allow successful startups to exit and technology to scale, authorities are increasingly vigilant regarding "non-traditional" mergers. The regulatory focus is on whether the merger effectively removes a viable independent competitor from the market that substantially lessens competition.

3.2 Summary of competition concerns raised by APAC regulators

As outlined above, regulators across the APAC region have flagged a wide range of potential competition concerns, although none have yet made adverse findings against specific firms. Table 4 below summarises the concerns identified by APAC regulators in recent reports.

229 NZCC, 'Navigating the rise of AI: Perspectives from a competition and consumer regulator', 7 July 2025, p 6.
230 ACCC, 'Digital platform services inquiry, final report', 13 March 2025, p 304.
NZCC, 'Navigating the rise of AI: Perspectives from a competition and consumer regulator', 7 July 2025, p 6.
231 NZCC, 'Navigating the rise of AI: Perspectives from a competition and consumer regulator', 7 July 2025, p 6.

Table 4: Theories of harm raised by APAC regulators

Theory of harm	ACCC	CCI	JFTC	KFTC	NZCC	TFTC
Input foreclosure	x		x	x		x
Self-preferencing	x	x	x		x	x
Tying and bundling	x	x	x	x	x	x
Ecosystem entrenchment	x	x	x	x		x
First mover advantage		x		x		
Market "tipping"	x	x			x	x
Strategic partnerships	x	x	x	x		x
Consumer inducement and exclusive dealing	x	x		x	x	x
Algorithmic collusion	x	x	x		x	x
Killer acquisitions	x	x			x	x
Talent attraction	x		x	x	x	x

Source: *RBB analysis based on desk research of publicly available sources involving a review of the following competition authority publications:*
*Australian Competition and Consumer Commission (**ACCC**), 'Digital platform services inquiry, final report', 13 March 2025*
*Competition Commission of India (**CCI**), 'Market study on artificial intelligence and competition', September 2025.*
*Japan Fair Trade Commission (**JFTC**), 'Report regarding Generative AI ver. 1.0', 6 June 2025*
*Korea Fair Trade Commission (**KFTC**), 'Generative AI and Competition', 18 December 2024.*
*New Zealand Commerce Commission (**NZCC**), 'Navigating the rise of AI: Perspectives from a competition and consumer regulator', 7 July 2025.*
*Taiwan Fair Trade Commission (**TFTC**), 'Competition Law Issues Related to Generative Artificial Intelligence, consultation paper', 7 September 2025.*

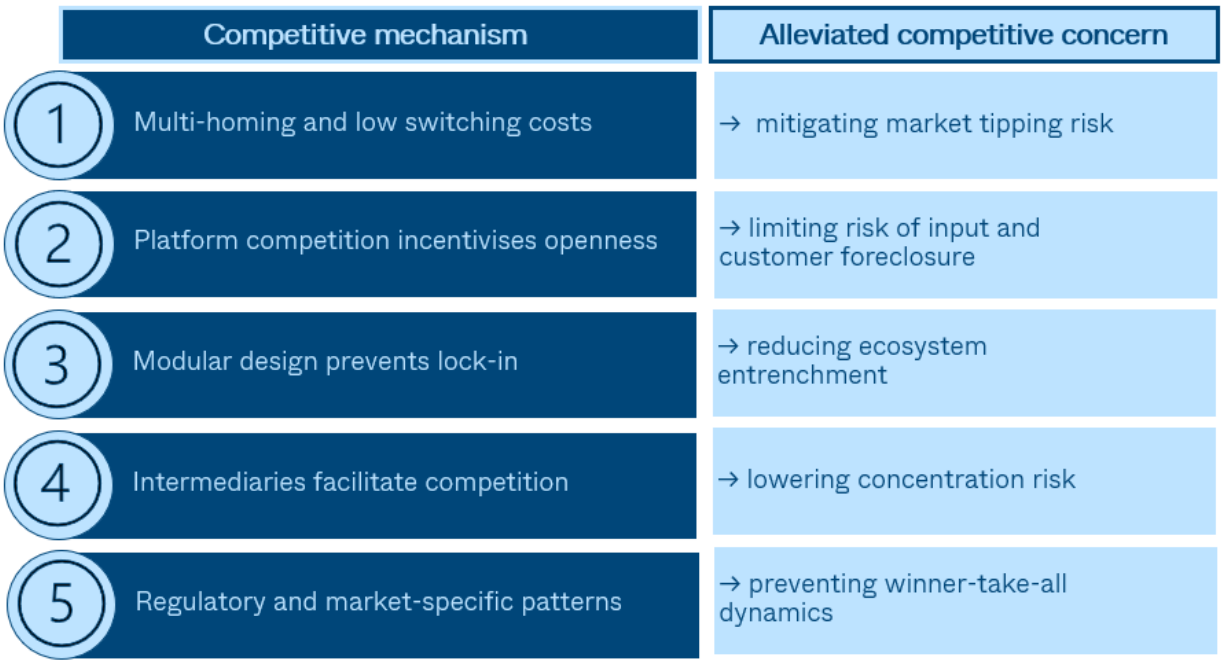
Notes: *Each authority uses different terminology and analytical frameworks when describing potential competition harms. As a result, the presence or absence of specific theories of harm is not always explicit. In several instances, conclusions rely on interpretation of the conduct described, the wider policy context, or reasonable inference where the authority touches on an issue without naming it directly.*
Customer foreclosure is omitted from the table as APAC regulators have not directly referred to customer foreclosure in their publications to date. As discussed above, there is some overlap between customer foreclosure and specific commercial strategies that may be used to achieve this such as tying, bundling, exclusive dealing, and self-preferencing, which are outlined by regulators and included in the table. The omission of direct references to customer foreclosure does not imply that foreclosure-type theories of harm are not being examined by APAC regulators.

4 Competitive implications for the Asia-Pacific region

Sections 2 and 3 outlined the APAC GenAI markets' distinctive characteristics and the potential competition concerns identified by competition authorities. This section examines whether these concerns have materialised or are likely to materialise, drawing on the deployment case studies detailed in section 5.

We identify five key competitive dynamics the APAC markets (summarised in Figure 9). To date, these mechanisms address the concerns outlined in section 3.

Figure 9: Overview of the five key competitive mechanisms currently shaping the GenAI market and addressing competition concerns



We summarise key points for each of these five competitive dynamics below.

- **Multi-homing and low switching costs (mitigate market tipping risks):** Organisations including super apps (Grab), conglomerates (Samsung), and financial institutions (CBA) actively use multiple FMs. Multi-homing serves two distinct purposes. First, specialisation: routing queries to whichever model performs best for that task (for example, Baidu using DeepSeek for advanced search). Second, substitution: maintaining redundant providers for the same task (for example, the Trae coding app offers five selectable models). This results in reduced lock-in concerns and easier switching between FM providers. This behaviour directly counters the risk of market tipping. If switching is feasible and low-cost (as demonstrated when Chinese developers migrated from OpenAI to domestic alternatives, most of which were open-source, within two weeks in mid-2024), it is difficult for any single provider to secure the lasting advantage required to tip the market.
- **Platform competition incentivises openness (limiting the risk of input and customer foreclosure):** Platforms compete to attract enterprise customers by offering choice. For example, Google Vertex AI distributes over 200 FMs from competing providers alongside Google's own Gemini models, while Amazon Bedrock offers over 100 third-party FMs alongside Amazon's Titan. Samsung includes competitor FMs within its enterprise platform despite developing its own proprietary models. This suggests that commercial incentives currently favour openness: restricting access would make the platform less attractive to business users. Consequently, this openness mitigates customer foreclosure concerns (rival FM developers can reach end users through these platforms) and input foreclosure

concerns (application developers can access competing FMs rather than being restricted to proprietary models).

- **Modular design prevents lock-in (reducing ecosystem entrenchment):** CBA maintains simultaneous partnerships with AWS, Microsoft, Anthropic and OpenAI, and design systems that can switch easily between providers and route tasks to the best-performing model. Orchestration tools such as OpenRouter or LangChain make supplier-neutral integration accessible to less sophisticated GenAI deployers. This approach treats FMs as interchangeable components, preventing "walled gardens" from forming and locking customers in.
- **Intermediaries facilitate competition (lowering concentration risk):** Aggregators and service integrators help users navigate the market, reducing the search costs that usually favour incumbents. For example, TCS built partnerships across all major platforms, using its WisdomNext tool to compare models and recommend optimal choices for each customer and task. Similarly, Samsung provides access to multiple FMs through its FabriX enterprise platform and helps customers compare models. By bridging different systems, these intermediaries make it easier for FMs to compete for business, helping counter potential market concentration at the model layer.
- **Regulatory and market-specific patterns (preventing "winner-take-all" dynamics):** Region-specific conditions in APAC prevent FM markets from tipping toward a single supplier. For example, Samsung uses Gemini globally but substitutes it with ERNIE in China where Google is blocked. Local language needs and strong mobile preferences can give domestic developers an advantage, as shown by Rakuten's Japanese-optimised models. These dynamics can lead to competition between global and local providers: global models adapt to local languages (as seen in Korea), while cost-efficient small language models (like Samsung's TRM) compete through on-device inference.

Sections 4.1-4.5 examine each dynamic in detail. Observed patterns of multi-homing behaviour and openness suggest competition is currently effective at both the development and deployment layers of the GenAI value chain.

4.1 Multi-homing patterns and low switching costs

The evidence we have observed suggests that switching costs between FMs remain low, and that customers frequently multi-home across FMs.

4.1.1 Multi-homing patterns, switching costs and market mechanisms enabling flexibility

As discussed in section 2.4.1.5, multi-homing across FM providers is widespread, as seen across different firm types and regions in APAC.

- **Grab** (section 5.2.2) deploys OpenAI, Anthropic and internally-developed models for different functions.
- **Samsung** (section 5.2.3) builds proprietary Gauss models, uses Google Gemini globally and Baidu ERNIE in China, commissions customer service tools from Microsoft Azure AI, and distributes third-party models through Samsung SDS FabriX.
- **TCS** (section 5.2.4) maintains partnerships across AWS, Google Cloud, Microsoft Azure and IBM watsonx, deploying different platforms for different clients.
- **Rakuten** (section 5.2.5) combines OpenAI for general-purpose capabilities with proprietary Japanese-optimised models.
- **CBA** (section 5.2.6) co-develops GenAI applications with AWS, Microsoft, Anthropic, and OpenAI.

- **Baidu** relies on DeepSeek for advanced search alongside in-house ERNIE.[232]
- **Trae**, a Singapore-based coding app, allows users to choose between built-in GenAI models from Google, Moonshot AI, OpenAI, DeepSeek and xAI.[233]
- **Wotton-Kearney**, an Australian legal firm, released a GenAI application tailored to legal workflows relying on FMs from Meta and Mistral AI.[234]

This multi-homing is driven by two distinct commercial needs, each impacting competition differently. In some cases, firms route different tasks to different FMs based on capabilities in that domain (for example, Baidu Search using DeepSeek for advanced queries alongside in-house ERNIE for standard queries). This maintains demand across multiple providers that may ease future switching. However, where a specific model is chosen for its unique capabilities, substituting an alternative for that particular task may not be straightforward. Albeit, innovation may mean that a model does not maintain a unique advantage over time. In other cases, firms maintain multiple FMs as substitutes for the same use case (for example, Trae offering users a choice of five FMs). This creates direct competitive pressure. Providers know their product can be swapped out for a rival's model, which disciplines pricing and quality. Both patterns sustain demand for multiple FM providers, though they affect switching costs differently.

Several market-level developments reduce barriers to switching FM providers. Below we discuss how open-source models, deployment platforms, orchestration tools, standardisation and multicloud infrastructure can all help to reduce switching costs.

Open-source adoption creates low-cost alternatives to proprietary offerings

As discussed in section 2.4.1.3, open-source FMs play a prominent role in APAC markets. Open-source models are free to download and use, so developers can test and deploy alternatives to proprietary models at low cost. The proliferation of open-source models from Chinese firms (including Qwen, ERNIE and DeepSeek) and their adoption across APAC illustrates this pattern.

When multiple open-source alternatives offer comparable capabilities to those of proprietary models, organisations can credibly threaten to switch, constraining proprietary providers' pricing power. As discussed in section 2.3.4, government involvement in AI development reinforces this dynamic, particularly in China where authorities view domestic open-source ecosystems as critical for technology independence (sections 2.4.1.3 and 5.1).

The persistent availability of open-source alternatives constrains proprietary providers. If prices rise or quality falls, developers can switch. This reduces concerns about market tipping toward a few proprietary suppliers by ensuring viable low-cost alternatives remain available even if proprietary markets consolidate.

Deployment platform aggregation enables comparison of competing models without separate supplier relationships

Deployment platforms host multiple competing FMs, both open-source and proprietary, making it easier to compare models and increasing competition among suppliers. Amazon Bedrock and Google Vertex AI provide unified API access to over 100 FMs from multiple suppliers through standardised interfaces.[235]

APAC-specific deployment platforms follow similar patterns. For instance:

---

232 TechNode, 'Baidu Search integrates DeepSeek and Large Model ERNIE for advanced search', 17 February 2025.
233 Trae Ide & Solo, 'Models', accessed November 2025.
234 Wotton Kearney, 'Wotton Kearney launches game-changing AI solution for Australian legal industry', 16 May 2024.
235 AWS, 'Amazon Bedrock', accessed November 2025.
Google Cloud, 'Vertex AI', accessed November 2025.

- **Alibaba Cloud Model Studio, Tencent's TI** and **Baidu's Qianfan** each offer access to competing Chinese and international models alongside their own proprietary FMs.[236]
- **Samsung SDS FabriX** (section 5.2.3), provides enterprise customers access to ChatGPT, Meta's Llama, DALL-E, Naver's HyperClova X and Samsung's proprietary models.[237]
- **TCS WisdomNext** (section 5.2.4), aggregates offerings from AWS, Google Cloud, Microsoft Azure and IBM watsonx, with "intelligent evaluator bots" that compare models to recommend optimal choices for specific workloads.[238]

These platforms offer different levels of support. Some route requests between providers for easy switching, while others provide evaluation tools, compliance features, and platform-specific optimisations. This variety allows organisations to select approaches matching their technical abilities and needs.

Deployment platforms enable enterprises to access FMs through a unified authentication, billing and security framework, reducing the need for separate supplier relationships. This also creates competitive pressure among model suppliers. Platforms place models side-by-side for a direct performance and cost comparison, and provide standardised interfaces that make switching straightforward. This means that FM providers must compete on price and quality to retain their customers.

Orchestration tools reduce barriers to multi-homing

Independent orchestration tools have emerged to reduce the technical complexity of managing multiple FM providers across different platforms. These tools operate as unified interfaces that account for differences between model providers behind the scene, enabling organisations to route across FM providers using standardised code.

OpenRouter provides access to over 500 models from over 60 providers through a single API endpoint that maintains compatibility with OpenAI's API format.[239] Developers can switch between providers by changing a single line of code (the API endpoint URL). The platform's traffic patterns indicate adoption across APAC markets, with India and China among significant user bases.[240] OpenRouter integrates Chinese models including Kimi (Moonshot AI's chatbot model), Qwen (Alibaba's model family) and DeepSeek.[241]

LangChain enables developers to write single sets of code that adapt for various FMs without requiring time-consuming rewrites for each provider.[242] The framework has achieved over 130 million total downloads globally and supports over 70 model providers, with survey evidence suggesting 33% of respondents that use open-source tools employ LangChain.[243] Similar orchestration platforms, including Portkey and LiteLLM, provide comparable multi-model access with additional enterprise features such as content filtering, usage monitoring and regulatory compliance tools.[244]

---

236 Alibaba Cloud, 'Alibaba Cloud for Generative AI', accessed November 2025.
Tencent Cloud, 'What mainstream large language models does the large model application building platform support?', 31 July 2025.
Baidu, 'Baidu Qianfan', accessed December 2025.

237 Samsung SDS, 'Samsung SDS Realizes Hyper-Automation Innovation with GPU-Centric AI Cloud', 4 September 2024.
Korea Economic Daily, 'Samsung SDS AI service FabriX to launch on Microsoft Azure platform', 3 September 2024.
Samsung SDS, 'Fabrix', accessed December 2025.

238 TCS, 'TCS AI WisdomNext™', accessed November 2025.

239 OpenRouter, 'About OpenRouter', accessed December 2025.
OpenRouter, 'Quickstart', accessed November 2025.
OpenRouter, 'API Reference', accessed December 2025.

240 Similarweb, 'Openrouter.ai Traffic Analytics', accessed November 2025.

241 OpenRouter, 'Models', accessed December 2025.

242 EM360Tech, 'What is LangChain and How to Use It', 18 June 2025.

243 LangChain, 'LangChain's Second Birthday', 24 October 2024.
McKinsey, 'Open source technology in the age of AI', 22 April 2025, p 5.

244 Portkey, 'Production Stack for Gen AI Builders', accessed November 2025.
TrueFoundry, 'Portkey vs LiteLLM : Which is Best ?', 4 April 2024.
Helicone, 'Top 5 LLM Gateways in 2025: The Complete Guide to Choosing the Best AI Gateway', 16 June 2025.

These orchestration tools actively enable multi-homing across FMs by offering unified access to multiple models and by including automatic fallbacks to alternate models. By requiring minimal technical adjustments to switch models, these tools can lower switching costs, especially for less sophisticated developers.

Standardisation initiatives accelerate interoperability and reduce switching friction

Industry standardisation efforts also make it easier to work with multiple providers simultaneously.

- Anthropic's MCP, introduced in November 2024 and subsequently endorsed by OpenAI and Microsoft, gained rapid adoption from major Chinese platforms.[245]
- In April 2025, Alibaba Cloud launched an MCP-compatible service marketplace with over 1,000 offerings and Baidu, Tencent and Alipay adopted MCP around the same time or shortly afterwards.[246]

This rapid uptake by Chinese platforms suggests the commercial value of shared technical standards. However, as discussed in section 2.4.3.4, MCP adoption remains geographically uneven within APAC. China's major platforms adopted MCP within five months and Western companies endorsed the protocol (OpenAI in March 2025, and Google DeepMind in April 2025).[247] Community-developed MCP servers exist for Naver APIs, but without official adoption.[248] This geographic variation may reflect different calculations on the benefits of interoperability versus proprietary control, or may simply indicate that standardisation has not yet reached sufficient maturity in all markets.

Multicloud infrastructure enables credible switching

Organisations in APAC can use multicloud setups to comply with regions varying regulatory requirements (discussed further in section 4.5.1).

This multicloud infrastructure has a competitive effect at both the platform and foundation model levels. Organisations using multiple cloud providers and their deployment platforms can directly test and compare FMs across a wider range than if limited to a single platform's offerings. Additionally, organisations can compare deployment platform offerings including tools for fine-tuning foundation models. As organisations can test and employ various FMs through multiple different platforms, this familiarity with alternatives can make the threat of switching credible.

4.1.2 Natural experiment: OpenAI China exit demonstrates rapid switching capability

OpenAI's July 2024 restriction of API access from mainland China and Hong Kong created a natural experiment testing organisations' ability to switch FM providers under severe time pressure. With only two weeks between announcement (24 June 2024) and enforcement (9 July 2024), developers building applications powered by OpenAI's models had to switch almost immediately.

More than ten Chinese providers announced migration programmes within days, offering substantial incentives in terms of data usage tokens and support designed to reduce switching costs.

- **Alibaba Cloud:** 22 million free tokens and migration services for app developers.[249]

---

245 Techzine, 'How the Model Context Protocol has taken the AI world by storm', 26 April 2025.
246 Techzine, 'How the Model Context Protocol has taken the AI world by storm', 26 April 2025.
SCMP, 'Chinese tech giants race to expand AI services market with MCP', 21 April 2025.
AIBase, 'Alibaba Announces Full Support for MCP Protocol', 9 April 2025.
247 TechCrunch, 'Google to embrace Anthropic's standard for connecting AI models to data', 9 April 2025.
248 While community-developed MCP servers exist for Naver APIs, major platforms including Naver, Line, and Southeast Asian super apps have not announced official MCP integration as of November 2025.
249 Yicai Global, 'Alibaba, Baidu, Other Chinese AI Firms Rush to Fill Gap Left by ChatGPT in China', 26 June 2024.

- **Zhipu AI:** "Special Migration Program" offering up to 150 million free tokens and training sessions.[250]

- **Baidu:** "Inclusive program" offering new users 50 million free tokens for ERNIE 3.5, specifically targeting existing OpenAI users with additional token packages equivalent to the scale of their OpenAI usage.[251]

- **Moonshot AI:** Developed proprietary APIs fully interoperable with OpenAI's API specifications, enabling migration processes as simple as replacing API endpoint URLs in application code.[252]

- **Minimax:** Offered one-click migration to its own models.[253]

Fine-tuning FMs on organisation-specific data creates potential switching costs, as customisations must be recreated when changing providers. However, providers recognised fine-tuning as a potential migration barrier and offered support. Baidu announced free fine-tuning for migrating organisations, while others offered migration services including assistance with fine-tuning workflows.[254] This demonstrates that while fine-tuning can create switching costs, providers competing for customers have incentives to reduce them.

Despite OpenAI's API restrictions, some access pathways remained available. Microsoft published a step-by-step migration guide to its local Azure service operated by partner 21Vianet, enabling developers to continue using OpenAI services through Microsoft's Azure infrastructure.[255] This provided an alternative switching path, with developers retaining their access to OpenAI FMs, but switching platform and infrastructure. Open-source FMs, such as Meta's Llama and localised versions such as Chinese-Llama-Alpaca, which has more Chinese vocabulary and uses local training data, remained available.[256]

As of early July 2024, there was limited disruption with applications continuing to operate and domestic providers gaining customers.[257] The rapid transition is consistent with organisations being able to switch providers even under tight time constraints. Detailed examination in section 5.2.1 shows that losing a major competitor triggered intense competition among remaining providers for former OpenAI customers via migration incentives and technical support.

4.1.3 Distinctive super app pattern: platform competition drives AI model competition

Super apps are more prevalent in APAC than other markets where users maintain fragmented relationships across specialised services. This creates distinct distribution dynamics for GenAI in the region.

WeChat (1.4 billion monthly active users), Alipay (1.4 billion), KakaoTalk (54 million), Paytm (270 million), LINE (224 million), Grab (46 million), and Gojek (38 million) all combine multiple services in single platforms.[258] Switching from one super app to another can impose costs: rebuilding social connections,

---

250 Reuters, 'Chinese AI firms woo OpenAI users as US company plans API restrictions', 26 June 2024.
Time, 'How OpenAI's Decision Not to Operate in China Will Reshape the Chinese AI Scene', 26 June 2024.
251 Baidu, 'Hometown Clouds: Domestic Large Model Universal Benefit Plan', 25 June 2024.
Reuters, 'Chinese AI firms woo OpenAI users as US company plans API restrictions', 26 June 2024.
252 Moonshot AI, 'Migrating from OpenAI to Kimi API', accessed November 2025.
253 China Daily, 'US firm's AI tech ban set to inspire homegrown innovation', 25 July 2024.
254 Baidu, 'Hometown Clouds: Domestic Large Model Universal Benefit Plan', 25 June 2024.
255 Time, 'How OpenAI's Decision Not to Operate in China Will Reshape the Chinese AI Scene', 26 June 2024.
256 Source Forge, 'Chinese-LLaMA-Alpaca-2 v2.0', accessed November 2025.
257 STCN, '"Relocation" or "Going Global": OpenAI's Countdown to the Discontinuation of its Chinese APIs', 2 July 2024.
258 This monthly active users for WeChat includes the users for Weixin (China) and WeChat (international users). Tencent, '2025 First Quarter Results', 14 May 2025.
CoinLaw, 'Alipay Statistics 2025: User Adoption, Transaction Volumes, and Technological Innovations', 16 June 2025.
Market.biz, 'Kakaotalk Statistics and Facts', 20 August 2025.
CoinLaw, 'Paytm Statistics 2025: Financial Performance and User Engagement Insights', 17 June 2025.

re-downloading mini apps, and re-verifying payment credentials. In addition, users may lose platform-specific data and customisations which can impart significant user value.

Despite potential lock-in concerns, super apps face competitive pressure from other rivals with similar offerings. WeChat competes with Alipay, Grab competes with Gojek, and LINE competes with KakaoTalk. Multi-homing across super apps is common, with 94% of internet users in China reporting using both WeChat and Alipay.[259] This suggests that setting up accounts on multiple super apps is straightforward enough that most users do so, undermining the lock-in concern. If users can easily maintain presence on competing platforms, no single super-app can capture them exclusively.

This competitive intensity is reflected by two key super app features:

- **Open mini app ecosystems.** Super app consumers value application diversity and quality. Super apps have therefore created broad and open mini app marketplaces that include a variety of services such as shopping platforms, mobile gaming, streaming services and ride-hailing from third-party developers. These mini app integrations even include applications from competing super apps. For example, WeChat and Alipay integrate each other's mini apps across their competing super apps: Didi ride-hailing is available in both ecosystems and Taobao, an e-commerce marketplace operated by Alibaba, is accessible through WeChat.[260] Competing Indian super apps Paytm and PhonePe both offer a similar suite of third-party mini apps, including Ola ride-hailing, which appears in both.[261]
- **FM multi-homing within core super app features.** GenAI features are emerging as important drivers of user engagement on super apps.[262] To provide attractive AI features and prevent customers switching to rival apps, super apps are integrating the best available GenAI FMs, regardless of provider. This multi-homing pattern reflects the matching of FMs with specific use cases as discussed in section 4.1.1. For instance:
  - Tencent (owner of WeChat/Weixin) tested integrating DeepSeek (an independent AI model) into Weixin search in February 2025 despite developing its own Hunyuan FM;[263]
  - Baidu (China's largest search engine) integrated DeepSeek alongside its ERNIE model in Baidu Search in February 2025 for advanced search queries;[264] and
  - Grab (section 5.2.2) actively deploys competing FMs within a unified super app user interface, selecting each provider based on specific capability requirements (OpenAI's GPT-4o for AI Driver Companion while using Anthropic's Claude for AI Merchant Assistant).[265]

This creates a distinctive APAC competitive pattern. For FM providers, super app integration offers distribution to large user bases. For example, when Tencent integrates DeepSeek into Weixin, DeepSeek reaches 1.4 billion users without building its own consumer product. For application developers, mini app ecosystems provide an established route to market. Both dynamics sustain competition by enabling

---

Digital Marketing for Asia, 'Why is LINE the most popular social media app in Japan?', accessed November 2025.
This figure is based on monthly transacting users. Grab, 'Grab Reports Second Quarter 2025 results', 31 July 2025.
Techwire Asia, 'Gojek sees profitability ahead after decade of rapid growth', 16 November 2020.

259 S&P Global, 'Mobile payments, mini-programs are key features of Chinese super apps', 20 October 2020.

260 SCMP, 'Chinese tech giants Tencent and Alibaba break digital wall with Taobao, WeChat integration', 9 October 2025.

261 PhonePe, 'PhonePe Partners with Ola to Launch the Industry-First AutoPay Feature', accessed November 2025.
Paytm, 'Paytm Mini App Store', accessed November 2025.

262 Market reports show that standalone AI apps and chatbot downloads grew at over 200% in 2024, while time spent on these apps grew at nearly 350% (both representing the highest growth rates among the top 20 app categories). While these growth rates come from a lower base than comparative app categories and relate to standalone apps rather than AI features within super apps specifically, they suggest AI features are generating user attention and could represent a sizeable opportunity for super apps. Further, apps mentioning AI represented 13% of all global app downloads in 2024, with commentary noting that, "*there is a growing trend of integrating AI-powered features into applications to attract tech-savvy users and enhance overall convenience*".
SensorTower, 'State of Mobile 2025', pp 12-14 and pp 24-25.

263 Reuters, 'Tencent's Weixin app, Baidu launch DeepSeek search testing', 17 February 2025.

264 TechNode, 'Baidu Search integrates DeepSeek and Large Model ERNIE for advanced search', 17 February 2025.

265 Grab, 'Grab deploys agentic AI to empower merchants and driver partners', 8 April 2025.

multiple providers to reach users, reducing the risk that any single FM or app becomes the default across the ecosystem due to super app openness.

4.2 Platform competition limits foreclosure incentives

Competition authorities have raised concerns that vertically integrated firms (those operating across multiple layers of the value chain) could restrict downstream rivals' access to upstream inputs (input foreclosure) or restrict upstream rivals' access to downstream customers (customer foreclosure). See section 3.1.1 for a fuller discussion.

Foreclosure is profitable only when the gains from weakening rivals exceed the costs. For input foreclosure, the gain is capturing downstream market share by weakening rivals denied access to the FM. The cost is foregone revenue from those downstream firms. For customer foreclosure, the gain is strengthening the FM by denying rivals access to customers. The cost is a less attractive platform that may lose customers to rivals offering more choice.

Current APAC conditions work against this. Platform competition is intense and constrains downstream market power (super apps compete on AI features), alternative FMs remain available through open-source and competing deployment platforms, and enterprise customers actively demand supplier flexibility.

Several APAC firms are vertically integrated and could hypothetically use these strategies. Yet observed behaviour demonstrates openness across different organisational structures and market contexts.

4.2.1 Vertical integration in the APAC GenAI value chain

Vertical integration takes different forms in APAC markets, across different layers of the GenAI value chain.

Chinese technology firms (section 5.1.1) operate cloud infrastructure, develop FMs and control major consumer platforms.

- **Baidu** operates Baidu AI Cloud, develops ERNIE FMs serving over 200 million users and owns China's leading search engine.[266]
- **Alibaba** operates Alibaba Cloud and Alibaba Cloud Model Studio, develops the Qwen open-source model family, and operates major e-commerce platforms through the Alipay super app.[267]
- **Tencent** operates Tencent Cloud, develops Hunyuan FMs, and operates WeChat/Weixin serving 1.4 billion users.[268]
- **ByteDance** operates ByteHouse cloud services, develops Doubao FMs, and owns the Douyin and TikTok platforms.[269]

[266] Baidu, 'Baidu Announces Fourth Quarter and Fiscal Year 2024 Results', 18 February 2025.
Reuters, 'Baidu says AI chatbot 'Ernie Bot' has attracted 200 million users', 16 April 2024.
Investopedia, 'Baidu vs. Google: Understanding China's Leading Search Engine', accessed November 2025.

[267] Communications Today, 'China's Q3 2024 cloud spending hits $10.2 billion'. 24 December 2024.
Business Wire, 'Alibaba Group Announces December Quarter 2024 Results', 20 February 2025.

[268] Communications Today, 'China's Q3 2024 cloud spending hits $10.2 billion'. 24 December 2024.
Tencent, 'Tencent Announces Global Rollout of Scenario-Based AI Capabilities to Accelerate Industrial Efficiency', 16 September 2025.
Mobile World Live, 'Tencent Q3 financials surge', 14 November 2025.

[269] Forbes, 'AI Mania Makes ByteDance Cofounder Zhang Yiming China's Richest Person', 9 March 2025.
Tech in Asia, 'ByteDance's Doubao stays China's top AI app as DeepSeek loses users', 18 September 2025.
ByteHouse, 'ByteHouse', accessed December 2025.

Samsung's chaebol structure (section 5.2.3) spans semiconductors, cloud services (Samsung SDS operating cloud and FabriX enterprise platform), FM development (Gauss models for internal use) and consumer device AI deployment (Galaxy AI deployed across 200 million devices by end 2024).[270]

Rakuten's integrated platform (section 5.2.5) combines e-commerce, fintech (Rakuten Card, Rakuten Bank and Rakuten Pay), telecommunications (Rakuten Mobile with 9 million subscribers), travel and digital content serving 1.7 billion members worldwide.[271] This integration generates cross-service proprietary data enabling development of specialised FMs and distribution advantages for services like semantic search (which aims to understand user intent), recommendations, and agentic AI launched in July 2025.[272]

Super app platforms integrate messaging, payments and multiple service categories while controlling distribution to hundreds of millions of users. Grab (section 5.2.2) operates across eight Southeast Asian countries and developed internal AI capabilities including Vision LLM built on Qwen2-VL 2B for document processing and Mystique AI for copywriting.[273]

4.2.2 Evidence of openness across vertically integrated firms

Evidence from APAC suggests that firms across diverse organisational structures favour openness.

Deployment platform operators distribute competitor models. Rather than restricting users to in-house offerings, platforms with proprietary models actively distribute rival FMs. Samsung SDS (section 5.2.3) operates FabriX which distributes Naver's HyperClova X alongside ChatGPT, Meta's Llama and DALL-E, despite Samsung developing its own Gauss FMs.[274] Chinese platforms (section 5.1) offer DeepSeek models alongside their proprietary models.

Super app openness predates GenAI. The commercial need to retain users appears to outweigh any incentive to attempt to lock rivals out. WeChat and Alipay both host third-party mini apps (Didi ride-hailing available in both ecosystems), and services from one ecosystem can be available through the rival: Taobao (Alibaba's e-commerce marketplace) is available as a mini-app in Tencent's WeChat.[275] This platform-level openness is an established pattern, now extending to GenAI. Platforms integrate rival AI FMs alongside their own to offer the best available features.

Integrated players maintain external partnerships. Rakuten (section 5.2.5) develops proprietary Japanese-optimised models (Rakuten AI, released as open-source under the Apache 2.0 licence in February 2025) and partners with OpenAI. Rakuten deploys OpenAI for general purpose capabilities and its enterprise product Rakuten AI for Business.[276] Samsung (section 5.2.3) partners with Google Gemini

---

[270] Samsung Newsroom, 'Samsung and Google Cloud Join Forces To Bring Generative AI to Samsung Galaxy S24 Series', 18 January 2024.
Samsung SDS, 'Samsung SDS Realizes Hyper-Automation Innovation with GPU-Centric AI Cloud', 4 September 2024.
Korea Economic Daily, 'Samsung SDS AI service FabriX to launch on Microsoft Azure platform', 3 September 2024.
ET News, 'Samsung Electronics selects MS as AI supplier for customer service', 2 September 2024.
TechRadar, 'Samsung's new Gauss 2 AI Model might be the next Galaxy brain', 22 November 2024.
BusinessWire, 'Samsung Expands Galaxy AI as Consumer Desire for Mobile AI Grows', 14 July 2025.

[271] Rakuten Mobile, 'Rakuten Mobile Surpasses 9 Million Subscribers', 7 July 2025.
Rakuten Symphony, 'Rakuten and VEON to Cooperate in Open RAN and Digital Services to Rebuild Ukraine's Infrastructure', 2 August 2023.

[272] Rakuten, 'Rakuten Announces Full-Scale Launch of Rakuten AI', 30 July 2025.

[273] The Register, 'LLMs are lousy at reading Asian languages, finds Singapore's Grab', 4 November 2025.
Grab, 'How we built a custom vision LLM to improve document processing at Grab', 4 November 2025.
Grab, 'Meet Mystique, the AI-powered copywriting tool that helps Grab communicate effectively', 6 September 2024.

[274] Korea Economic Daily, 'Samsung SDS AI service FabriX to launch on Microsoft Azure platform', 3 September 2024.
Samsung SDS, 'Fabrix', accessed December 2025.
Samsung SDS, 'Samsung SDS Realizes Hyper-Automation Innovation with GPU-Centric AI Cloud', 4 September 2024.

[275] SCMP, 'Chinese tech giants Tencent and Alibaba break digital wall with Taobao, WeChat integration', 9 October 2025.

[276] Rakuten, 'Rakuten to Collaborate with OpenAI to Develop State-of-the-Art AI Experiences for Consumers and Businesses', 2 August 2023.
Rakuten, 'Rakuten Selects OpenAI as Strategic AI Partner and Collaborator on New "Rakuten AI for Business" Platform', 14 November 2023.
Rakuten, 'Rakuten Releases High-Performance Open Large Language Models Optimised for the Japanese Language', 21 March 2024.
Rakuten, 'Rakuten Unveils New AI Models Optimised for Japanese', 18 December 2024.

globally (announced January 2024, and extended to 200 million devices by the end of 2024), Baidu ERNIE in China for consumer devices (announced January 2024), Microsoft Azure AI for customer service (selected in September 2024 following a competitive evaluation), and develops internal Gauss models that are deployed to over 60% of internal developers.[277] The Samsung example is particularly salient as the company's activities span from AI semiconductor design to consumer electronics, which could hypothetically give it the commercial and technical ability to create a closed, self-sufficient ecosystem, or "walled garden". Despite this, Samsung maintains openness and partners with providers across the value chain.

4.3 Design choices enable partnerships without lock-in

Organisations can maintain flexibility by designing internal platforms that sit between applications and external suppliers, so switching providers does not require rewriting systems.

CBA's co-development with four major FM providers (section 5.2.6) demonstrates that firms can form both deep and broad partnerships. Samsung's announcement of a Microsoft Azure AI partnership in September 2024, eight months after its Google Cloud partnership, further shows that co-development with one supplier does not preclude partnership with others (section 5.2.3).

4.3.1 Collaborations with multiple suppliers are a feature of the deployment layer

CBA (section 5.2.6) operates co-development relationships with AWS, Microsoft, Anthropic and OpenAI.

- **The AWS partnership** evolved over a decade. Cloud migration began in 2015, the firms opened the AI Factory in September 2024, undertook full data migration in June 2025, and started a five-year strategic collaboration with AWS as a "preferred cloud provider" in February 2025.[278]
- **The Microsoft partnership** (announced in March 2024) involves Microsoft engineers from Seattle headquarters working directly with CBA teams, deploying Microsoft Copilot to 10,000 employees and co-developing CommBank Copilot.[279]
- **The Anthropic partnership** (expanded in March 2025) involves CBA engineers working directly with Anthropic's AI experts on fraud prevention and customer service.[280] Although CBA has not released the details of the models used, CBA has achieved measurable fraud detection results using GenAI: processing over 20 million daily payments, sending 20,000-35,000 daily alerts and contributing to a 30% fraud reduction.[281]
- **The OpenAI partnership** (announced in August 2025) designates CBA as OpenAI's strategic banking partner and focuses on fraud detection.[282]

Rakuten, 'Rakuten AI 2.0 Large Language Model and Small Language Model Optimised for Japanese Now Available', 12 February 2025.

277 Samsung Newsroom, 'Samsung and Google Cloud Join Forces To Bring Generative AI to Samsung Galaxy S24 Series', 18 January 2024.
Techwire Asia, 'Baidu AI Cloud partners Samsung Electronics China on generative AI smartphone', 29 January 2024.
Bloomberg, 'Samsung to Showcase Baidu's Ernie AI in Latest Galaxy Phones', 29 January 2024.
Samsung SDS, 'Samsung SDS Realizes Hyper-Automation Innovation with GPU-Centric AI Cloud', 4 September 2024.
Korea Economic Daily, 'Samsung SDS AI service FabriX to launch on Microsoft Azure platform', 3 September 2024.
ET News, 'Samsung Electronics selects MS as AI supplier for customer service', 2 September 2024.
TechRadar, 'Samsung's new Gauss 2 AI Model might be the next Galaxy brain', 22 November 2024.
BusinessWire, 'Samsung Expands Galaxy AI as Consumer Desire for Mobile AI Grows', 14 July 2025.

278 CBA, 'CommBank revolutionises banking by activating AI Factory with AWS', 17 September 2024.
CBA, 'CommBank accelerates AI integration with major data migration to cloud', 4 June 2025.
CBA, 'CommBank and AWS expand collaboration to deliver global best cloud and AI capabilities, enabling idea to production in six weeks', 4 February 2025.
AWS, 'How CommBank made their CommSec trading platform highly available and operationally resilient', 19 August 2025.

279 CBA, 'CBA and Microsoft deepens Gen AI partnership', 11 March 2024.
Microsoft, 'Commonwealth Bank of Australia invests in AI skills and Microsoft Copilot', 23 June 2025.

280 CBA, 'CommBank expands strategic partnership with Anthropic', 14 March 2025.

281 CBA, 'Customer safety, convenience and recognition boosted by early implementation of Gen AI', 28 November 2024.

282 CBA, 'CommBank and OpenAI embark on Australia-first strategic partnership to advance AI solutions ', 13 August 2025.

These relationships involve joint engineering and co-created solutions. CBA established a dedicated Technology Hub in Seattle in March 2025, positioning engineers close to its partners' headquarters (AWS and Microsoft are Seattle-based, and OpenAI and Anthropic have Seattle offices). CBA technologists have participated in three-week exchange programmes working alongside teams from AWS, Microsoft, Anthropic, H2O.ai and OpenAI.[283]

4.3.2 Design choices preserve flexibility to use different suppliers

CBA (section 5.2.6) stays flexible by building general systems that are separate from a specific implementation for any one supplier. The bank's Development Operations Hosting Platform, enhanced in 2024-2025, provides standardised deployment infrastructure for AI applications.[284] Within this hosting platform, CBA has integrated Amazon Bedrock and Amazon Bedrock Knowledge Bases which provide access to over a hundred FMs from leading AI companies.[285] This platform enables CBA to streamline and standardise its internal AI application development and maintain choice across multiple models.

CBA has implemented "out of the box" guardrails within its hosting platform which enforce CBA's security and compliance policies on all model inputs and outputs to protect against malicious inputs and misleading output.[286] These guardrails are pre-configured which means CBA can swiftly develop safe and compliant AI applications regardless of the underlying FM used.

When CBA co-developed the CommBiz GenAI messaging service with AWS, the solution provided access to multiple models including Claude 3 (Anthropic) and Cohere LLMs.[287] CBA's approach shows that sophisticated buyers can develop infrastructure that enables deep integration with multiple providers.

4.3.3 Parallel development relationships reflect active supplier evaluation

Observed behaviour shows organisations successfully develop supplier-neutral designs and actively evaluate alternatives, which is likely to maintain competitive pressure at the FM level.

CBA (section 5.2.6) has both Anthropic and OpenAI partnerships for overlapping fraud and scam detection use cases. The Anthropic partnership (expanded in March 2025) involves developing scam and fraud protection capabilities.[288] The OpenAI partnership (announced August 2025, five months after the Anthropic expansion) involves parallel work on scam and fraud protection.[289] This multi-homing for the same safety-critical use case could constrain FM pricing by reducing switching costs and increasing the credibility that CBA can switch to a competing FM.

Samsung (section 5.2.3) demonstrates the same pattern. Despite already partnering with Google Cloud, Samsung approached multiple suppliers as prospective partners for a GenAI customer service application. Samsung conducted an internal technical and competitive evaluation of Microsoft, Google and Palantir solutions during July 2024, before selecting Microsoft Azure AI (which achieved the highest performance scores) in September 2024.[290] This suggests that companies view leading FMs as close substitutes for at least some use cases. They are willing to test multiple options for the same task to ensure they select the right supplier for optimal performance. This case further shows that co-development with one supplier

283 CBA, 'CommBank establishes Seattle Tech Hub', 27 March 2025.
284 CBA, 'CommBank and AWS expand collaboration', 4 February 2025.
285 CBA, 'CommBank and AWS expand collaboration', 4 February 2025.
AWS, 'Amazon Bedrock', accessed November 2025.
286 CBA, 'CommBank and AWS expand collaboration', 4 February 2025.
CBA, 'Customer safety, convenience and recognition boosted by early implementation of Gen AI', 28 November 2024.
287 CBA, 'CommBank and AWS expand collaboration', 4 February 2025.
288 CBA, 'CommBank expands strategic partnership with generative AI company, Anthropic', 14 March 2025.
289 CBA, 'CommBank and OpenAI embark on Australia-first strategic partnership to advance AI solutions ', 13 August 2025
290 Electronic Times, 'Samsung Electronics selects Microsoft as AI supplier for customer service', 2 September 2024.

(Google Cloud) does not prevent competitive evaluation and partnership with different suppliers (Microsoft Azure) for different tasks, as deployers choose the best model for specific use cases.

4.4 Intermediaries and platform aggregators facilitate competition

Service integrators and platform aggregators actively facilitate competition between FM providers by reducing deployment costs, and maintaining neutral access to multiple providers.

4.4.1 Service integrators facilitate GenAI deployment

When integrators can credibly deploy solutions on any major platform and compare providers objectively, platforms must compete on pricing, capabilities and integration support. This creates competitive pressure at the FM layer even from enterprises with limited direct technical capability.

TCS (section 5.2.4), India's largest IT services firm, has built partnerships with AWS, Google Cloud, Microsoft Azure, IBM watsonx and Nvidia.[291] The company trained 25,000 associates on Microsoft Azure OpenAI and upskilled and certified 25,000 associates on AWS generative AI services. This indicates a commitment expertise across platforms and FMs, enabling TCS to recommend and deploy whichever provider best suits each client.[292]

Client deployments reflect client-specific recommendations:

- **FairPrice Group** (one of Singapore's largest retailers): Google Cloud's Vertex AI, BigQuery, and Gemini models.[293]
- **Wyndham Hotels & Resorts** (9,100 hotels across 95 countries): AWS-based solutions.[294]
- **Manufacturing clients**: Microsoft Azure integration.[295]

Integrators provide the missing expertise that non-technical businesses need to deploy AI effectively, expanding the total market for FM providers by reaching customers who would otherwise find deployment too difficult. Intermediaries can also improve output accuracy through techniques like Retrieval-Augmented Generation that involves referring to an authoritative knowledge base, and prompt engineering to more effectively guide the AI, reducing error rates and enabling GenAI integration into business-critical functions.[296]

Further, intermediaries can reduce the risk of market tipping by lowering enterprise search costs while maintaining model choice. Rather than researching dozens of competing models independently, enterprises can rely on integrators for comparison frameworks and recommendations. As discussed in section 4.1.1, TCS's WisdomNext platform acts as a neutral broker.. Because its business depends on finding the best solution for each client, it has incentives to recommend the best model rather than promote specific suppliers.

---

291 TCS, 'TCS Partners with Google Cloud to Integrate Gemini Enterprise for its Workforce and Customers', 14 October 2025.
TCS, 'TCS to Build New AI-Led Solutions for Business Transformation in Collaboration with Microsoft', 20 June 2025.
TCS, 'TCS Launches New Generative AI Practice in Partnership with AWS', 27 November 2023.
IBM, 'IBM Launches watsonx Code Assistant, Delivers Generative AI-powered Code Generation Capabilities Built for Enterprise Application Modernization', 26 October 2023.
TCS, 'TCS Launches NVIDIA Business Unit to Accelerate AI Adoption for Customers Across Industries', 24 October 2024.

292 TCS, 'TCS Bets Big on Azure Open AI', 6 July 2023.
TCS, 'TCS Launches New Generative AI Practice in Partnership with AWS', 27 November 2023.

293 FairPrice Group, 'FairPrice Group opens Store of Tomorrow at Punggol Digital District', 28 August 2025.
Google Cloud, 'FairPrice Group Unveils 'store of Tomorrow' Program with Google Cloud to Reimagine Its Retail Experiences and Operations', 5 June 2025.

294 TCS, 'TCS Launches New Generative AI Practice in Partnership with AWS', 27 November 2023.

295 TCS, 'TCS Launches 5G-Enabled Cognitive Plant Operations Adviser to Help Transform Plant Operations', 14 March 2023.

296 Shuster et al., find that RAG pipelines can reduce hallucination rates by over 60%, with improved knowledgeability scores on both in-scope and out-of-scope domains compared to non-RAG models.
Shuster et al., 'Retrieval Augmentation Reduces Hallucination in Conversation', 2021, p 2.

Platform providers actively compete for these integrator relationships. TCS has co-developed GenAI solutions with AWS, Google and Microsoft, suggesting platforms view integrators as important channels.

4.4.2 Orchestration tools make it easier to access, compare and switch between models

As discussed in section 4.1.1, independent orchestration tools such as LangChain and OpenRouter reduce the technical complexity of managing multiple FM providers and help deployers multi-home across different FM providers. Furthermore, these tools provide features that (i) evaluate FM performance, and (ii) ease switching between different FMs, without wholesale changes to underlying code.

OpenRouter's Auto Router selects the best model for each specific query based on prompt complexity, task type and model capabilities.[297] Users can set preferences that dictate how the tool selects models by ranking preferred models or prioritising different attributes such as price and latency.[298] This automates model evaluation and switching, and can provide competitive pressure that constrains FM provider pricing and maintains quality.

LangChain also provides a model evaluation tool through its LangSmith platform which allows developers to track model performance through metrics such as cost, latency and response quality.[299] The LangSmith evaluation framework provides multiple modes through which to assess outputs (human evaluation, coded rules and LLM evaluation), as well as providing direct pairwise comparisons of outputs.[300] It makes it easy to compare performance across different foundation models, so developers can quickly switch to the model that best meets their needs.

4.5 Regulatory and market-specific patterns prevent winner-takes-all dynamics

APAC markets demonstrate competitive dynamics shaped by regulatory diversity, linguistic requirements, super app ecosystems and distinctive innovation approaches. These patterns do not merely segment markets between different providers; they actively support competition by driving infrastructure choices that reduce lock-in, creating demand for locally-optimised models, and sustaining multiple viable competitors in the APAC region at both the FM and application deployment level.

4.5.1 Regulatory diversity supports multi-provider strategies

As discussed in section 2.3.1, data sovereignty requirements across APAC require local data storage while permitting cross-border transfers under specified conditions. Indonesia's GR 71/2019 requires public electronic system operators to store data locally, while private operators may store data offshore subject to government access for supervision.[301] Vietnam's Decree 53/2022 requires certain foreign service providers to store user data locally and establish a local presence upon a government request.[302] Singapore's PDPA restricts cross-border transfers of personal data unless equivalent protections are implemented by the foreign entity.[303] China's cybersecurity and data security laws requires critical information infrastructure operators to store personal data locally, with foreign cloud providers accessing the market through local partners (such as, Microsoft Azure operating through 21Vianet as a separate entity in China).[304]

---

297 OpenRouter, 'Auto Router', accessed December 2025.
298 OpenRouter, 'Provider Routing', accessed December 2025.
299 LangChain, 'Know what your agents are really doing', accessed December 2025.
300 LangChain, 'Evaluation concepts', accessed December 2025.
301 Global Compliance News, 'Indonesia: New Regulation on Electronic Systems and Transactions', 7 November 2019.
302 DLA Piper, 'Vietnam: Cybersecurity regulations for data storage and setting up a branch office', 18 October 2022.
303 Singapore Statutes Online, 'Personal Data Protection Act 2012', accessed December 2025, part 6.
304 Chambers and Partners, 'Cloud Computing 2025 – China', 7 October 2025.
Microsoft, 'Microsoft Azure operated by 21Vianet', accessed December 2025.

These requirements drive multicloud adoption. IDC finds that almost 90% of Asia-Pacific enterprises now have meaningful workload deployments on multiple public clouds.[305] Each cloud platform offers access to different FM sets: AWS Bedrock provides access to Claude, Llama and Mistral; Google Vertex AI to Gemini and Claude; Alibaba Cloud Model Studio to Qwen, DeepSeek and Kimi.[306] Organisations using multiple cloud providers and their deployment platforms can test and compare FMs across a wider range than if limited to a single platform's offerings, building familiarity with alternatives that makes switching easier.

Regulatory restrictions also prevent any single FM achieving region-wide dominance. Samsung (section 5.2.3) partners with Google Gemini for Galaxy AI features globally but substitutes Baidu's ERNIE Bot in China where Google services are blocked, serving different markets with different providers.

Regulatory fragmentation creates genuine costs alongside these competitive benefits. As discussed in section 2.3.1, fragmented data rules increase costs and complexity. Firms cannot simply deploy a single strategy across the entire region as they might across EU member states under harmonised frameworks. Instead, they face compliance costs from navigating different data residency requirements in each jurisdiction and technical costs from maintaining multiple model integrations or regional variants. This can create barriers to entry for smaller entrants that lack sophisticated compliance capabilities.

Nevertheless, these dynamics limit the risk of market tipping. Multicloud adoption helps maintain competitive pressure within markets, while regulatory restrictions can increase the number of suppliers operating across APAC.

4.5.2 Local language requirements and mobile-first applications

APAC's linguistic diversity creates natural competitive advantages for local model developers. As discussed in section 2.4.1.4, the region encompasses languages with fundamentally different characteristics to Indo-European language families, requiring distinct technical approaches.

APAC societies are typically linguistically diverse, mobile-first, and display varying literacy rates. These features create distinctive requirements for GenAI deployment. Voice-based queries are prevalent. Over one-third of Google Search queries in India are voice-based, compared to 5% in developed markets.[307] Varying literacy rates and linguistic diversity favour voice interfaces over text-only interaction. This requires models to incorporate text, voice and image formats, creating opportunities for domestic developers to build fine-tuned multimodal models handling local languages and dialects.

Technical optimisation for language-specific characteristics creates measurable advantages.

- **Rakuten** (section 5.2.5) developed a custom tokeniser algorithm optimised for Japanese, representing text in fewer, larger chunks and reducing processing costs compared to English-based models that fragment characters.[308] Rakuten AI 7B, released in March 2024, achieved the best performance among open Japanese LLMs at launch.[309] Yet, Rakuten Mobile's integrated voice assistant combines Rakuten AI for Japanese language understanding with OpenAI's Realtime API for speech-to-speech capabilities, illustrating how applications can split tasks between local and global models rather than relying solely on one language-optimised model.[310]

---

305 IDC, 'Asia/Pacific State of Cloud: Adoption Trends, Challenges, and Preferences', 21 October 2024.
306 Google, 'Vertex AI Platform', accessed December 2025.
AWS, 'Model choice', accessed December 2025.
Alibaba Cloud Model Studio 'Model invocation', accessed December 2025.
307 Ernst & Young, 'The Aldea of India 2025: How much productivity can GenAI unlock in India?', 14 January 2025, p 25.
308 Rakuten Today, 'Rakuten's Open LLM Tops Performance Charts in Japanese', 20 April 2024.
309 Rakuten, 'Rakuten Releases High-Performance Open Large Language Models Optimised for the Japanese Language', 21 March 2024.
310 Rakuten Today, 'Rakuten Mobile collaborates with OpenAI to integrate new technology', 23 October 2025.

- **Naver** (South Korea) trained HyperClova X on 6,500 times more Korean data than GPT-4, demonstrating superior performance on Korean language benchmarks against competing models including Falcon, SOLAR and Llama.[311] However, new models such as OpenAI's o1, GPT-4o and Anthropic's Claude 3.5 Sonnet outperformed HyperClova X on the Korean College Scholastic Ability Test, indicating leading multilingual models from global developers may be closing the gap in Korean language understanding and may hold advantages in more complex reasoning tasks.[312]
- **Grab** (section 5.2.2) built Vision LLM on Qwen2-VL 2B, selected specifically for its efficient Southeast Asian language tokeniser supporting Thai and Vietnamese.[313] The custom model achieves enhanced performance processing Southeast Asian identity documents compared to generic English-optimised models.[314] Grab also deploys voice features customised to Southeast Asian languages for driver and consumer applications, including enabling visually impaired users to book rides using voice prompts.[315] [316]
- **Gojek's** AI assistant Dira helps customers navigate GoPay banking app functions through voice commands, powered by Google Cloud's Gemini 1.5 Flash.[317] This feature is designed to work on limited-capacity mobile phones and is reported to be the first voice assistant of this kind to understand Bahasa Indonesian.[318]

Language-specific optimisation can give local developers an advantage. Global providers offer broad multilingual support but may struggle to match the performance of domestic models trained on region-specific data. However, this advantage has limits.

First, applications can separate user-facing language handling from underlying processing. A local model might manage interaction in Korean or Malay while other models perform the core reasoning or analysis, as seen in Rakuten's combination of its own Japanese model with OpenAI's voice capabilities. This limits the local language advantage to one component of the application rather than the whole system. Second, global developers can invest in local language capabilities when market opportunity justifies it. The Naver example illustrates this: HyperClova X's Korean language advantage is narrowing as global models improve their multilingual performance.

The result is that FM markets in APAC can support multiple providers without consolidating around either a few global firms with powerful general models or local specialists with strong positions in their language niches. Specialisation can help foster competition between global and local firms.

4.5.3 Innovation in ultra-efficient SLMs and cost-efficient training techniques

Research in APAC shows that firms can compete through efficiency and smart design, not just scale.

Samsung's Advanced Institute of Technology (**SAIT**) AI Lab in Montreal developed the Tiny Recursion Model (**TRM**), a 7-million-parameter model that outperforms models 10,000 times larger on specific reasoning benchmarks with a claimed training cost of less than USD 500 (examined in detail in section

311 The Korea Herald, 'Korea's AI challengers take on ChatGPT with own LLMs', 1 September 2025.
Naver Cloud, 'HyperClova X Technical Report', 13 April 2024, Figure 1.
312 GitHub, 'Korean SAT LLM Leaderboard', accessed November 2025.
313 Grab, 'How we built a custom vision LLM to improve document processing at Grab', 4 November 2025.
The Register, 'LLMs are lousy at reading Asian languages, finds Singapore's Grab', 4 November 2025.
314 Ibid.
315 Vulcan Post, 'Grab launches its first AI Centre of Excellence, plans to hire at least 50 "high-value" roles', 23 May 2025.
316 Grab, 'Grab deploys agentic AI to empower merchants and driver partners', 8 April 2025.
317 PR newswire, 'GoTo Group and Google Cloud Extend Collaboration on Generative AI with Groundbreaking In-app Voice Assistant, Dira', 24 September 2024
318 Tech in Asia, 'GoTo launches new AI strategy with the introduction of Dira, the first ever AI-based fintech Voice Assistant in Bahasa Indonesia', 16 July 2024.

5.2.3.2).[319] Released as open-source, TRM uses recursive reasoning approach. Rather than training a larger model with more parameters, TRM applies the same small model repeatedly, refining its answer through multiple passes. This demonstrates that small, highly-targeted models can achieve excellent results on narrow reasoning tasks without large capital investment.[320]

TRM demonstrates several competitive dynamics. The low training cost and open-source release reduce barriers to entry for specialised reasoning tasks. For mobile-first APAC markets, on-device deployment offers advantages: applications work offline in areas with poor connectivity, respond faster without server round-trips, and keep data local, which addresses data sovereignty concerns. Ultra-efficient models like TRM can run on mobile hardware, creating competitive pathways that bypass cloud infrastructure providers entirely. Notably, TRM emerged from Samsung, a firm with sufficient resources to develop large models and to pursue frontier model partnerships (with Gemini and ERNIE) for consumer products.[321] This suggests efficiency-focused innovation is a strategic choice, not merely a fallback for resource-constrained developers.

DeepSeek's releases illustrate similar efficiency innovation, potentially driven by export restrictions on advanced semiconductors to China. DeepSeek-V3 reportedly cost only USD 5.58 million to train compared to over USD 100 million for OpenAI's GPT-4, while achieving comparable benchmark scores.[322] Additionally, DeepSeek's models are open-source, allowing other firms to build on them.

[319] VentureBeat, 'Samsung AI researcher's new, open reasoning model TRM outperforms models 10,000X larger — on specific problems', 8 October 2025.
AIM, 'Tiny Model from Samsung AI Lab Beats Gemini 2.5 Pro, o3-mini on ARC-AGI', 9 October 2025.
Jolicoeur-Martineau, Alexia, accessed November 2025.

[320] VentureBeat, 'Samsung AI researcher's new, open reasoning model TRM outperforms models 10,000X larger — on specific problems', 8 October 2025.
GitHub, 'Samsung SAIL Montreal / Tiny Recursive Models', accessed November 2025.
SiliconANGLE, 'Samsung researchers create tiny AI model that shames the biggest LLMs in reasoning puzzles', 9 October 2025.

[321] Techwire Asia, 'Baidu AI Cloud partners Samsung Electronics China on generative AI smartphone', 29 January 2024.
Bloomberg, 'Samsung to Showcase Baidu's Ernie AI in Latest Galaxy Phones', 29 January 2024.

[322] DeepSeek, 'DeepSeek-V3 Technical Report', 27 December 2024.
Wall Street Journal, 'The Next Great Leap in AI Is Behind Schedule and Crazy Expensive', 20 December 2024.

5 APAC case studies

This section summarises each case study expanding on the evidence presented in section 4. This section does not discuss the observed competitive dynamics in detail and instead describes the competitive landscape from which these dynamics emerge.

Section 5.1 examines China's FM landscape, which receives standalone treatment due to the market size, regulatory distinctiveness, and the unique competitive dynamics associated with the "war of a hundred models". Sections 5.2.1 to 5.2.6 present six detailed case studies spanning different market contexts and firms (a super app, conglomerates, a service integrator and a bank).

5.1 China's competitive FM landscape

We analyse China's FM landscape before the six deployment case studies for three reasons.

1. Market scale: China's projected AI capital expenditure for 2025 (USD 84-98 billion) substantially exceeds other APAC markets, with numerous FMs including ERNIE, DeepSeek, Qwen, Hunyuan and Doubao.[323]
2. Structural distinctiveness: China's closed ecosystem creates competitive dynamics not replicated in other APAC markets where global platforms compete directly with local providers.
3. The "war of a hundred models" involved intense price competition and rapid switching (following OpenAI's exit) which may illuminate switching abilities that are present in APAC more broadly.

5.1.1 Major players in China's GenAI market

Established technology firms active in search, e-commerce, and social media have invested heavily in GenAI FM development.

- **Baidu**, China's largest search engine, has developed its ERNIE FMs that are optimised for the Chinese language and serve over 200 million users through their downstream ERNIE Bot application.[324]
- **Alibaba**, one of the world's largest e-commerce companies, has developed its Qwen series of open-source LLMs with over 600 million downloads supporting 119 languages and dialects.[325]
- **Tencent**, owner of WeChat/Weixin, has developed its Hunyuan series of LLMs, and has released over 30 models in the past year including its Hunyuan 3D series models which have been downloaded over 2.6 million times.[326]
- **ByteDance**, owner of TikTok/Douyin, has developed its own LLM which powers its popular downstream Doubao chatbot with 157 million monthly users.[327]
- **Xiaomi,** a consumer electronics and smartphone manufacturer, recently released its first FM which beat models from OpenAI and Alibaba on a range of mathematics and coding tasks.[328]

---

323 TechWire Asia, 'China to deploy $98bn in AI investment this year amid US tech rivalry', 26 June 2025.
TechNode, 'China approves 346 generative AI services under national registration scheme', 9 April 2025.
324 Baidu Research, 'ERNIE Bot: Baidu's Knowledge-Enhanced Large Language Model Built on Full AI Stack Technology', 24 March 2023.
Reuters, 'Baidu says AI chatbot 'Ernie Bot' has attracted 200 million users', 16 April 2024.
Investopedia, 'Baidu vs. Google: Understanding China's Leading Search Engine', accessed November 2025.
325 Alizila, 'Alibaba Recognized on Fortune's 2025 Change the World List for Open-Source AI', 25 September 2025.
326 Tencent, 'Tencent Announces Global Rollout of Scenario-Based AI Capabilities to Accelerate Industrial Efficiency', 16 September 2025.
327 Forbes, 'AI Mania Makes ByteDance Cofounder Zhang Yiming China's Richest Person', 9 March 2025.
Tech in Asia, 'ByteDance's Doubao stays China's top AI app as DeepSeek loses users', 18 September 2025.
328 Dao Insights, 'Xiaomi launches its first open-sourced reasoning LLM', 7 May 2025.

The majority of these tech firms are active at multiple levels of the GenAI value chain as shown in Figure 10.

Figure 10: Notable GenAI products from China's technology firms[329]

Company	Notable GenAI models/platforms/apps
Baidu 百度	• Qianfan platform • ERNIE foundation models • ERNIE Bot app
Alibaba	• Alibaba Cloud Model Studio platform • Qwen foundation models • Quark AI assistant app
Tencent 腾讯	• Tencent Cloud TI platform • Hunyuan foundation models • Yuanbao AI assistant app
ByteDance	• BytePlus ModelArk platform • Doubao foundation models • Doubao AI chatbot app

Source: RBB summary based on desk research of publicly available sources.

These established technology companies have also invested in and collaborated with a range of AI startups that are developing their own models, encouraging further innovation and sustained competition. Notably, China's "six little tigers", Zhipu, MiniMax, Moonshot, Baichuan, 01.AI, and StepFun, which all have valuations of over USD 1 billion, have played a significant role in China's broader FM development.[330]

- **Zhipu**, founded in June 2019, released the open-source GLM model family and offers a range of products including ChatGLM (conversational chatbot), AutoGLM (personal assistant), and Ying (a text-to-video tool). Zhipu's GLM is notable for handling both Chinese and English efficiently, combining training techniques used by Google's BERT and OpenAI's GPT.[331]
- **MiniMax**, founded in early 2022, has multiple flagship GenAI models across different formats including its M2, Speech 2.6, Music 2.0, and Hailuo 2.3 models which power its various applications including Agent (chatbot), Video Hailuo (video creator), Audio (audio generator), and Talkie / Xingye (a character generator and chatbot). MiniMax's models and products have served over 212 million users and over 100,000 enterprises and developers to date.[332]
- **Moonshot**, founded in March 2023, developed the Kimi family of FMs. These models power Moonshot's customer-facing Kimi chatbot, which can process up to 2 million Chinese characters in one prompt. This makes it particularly effective for tasks involving long documents.[333] Moonshot continues to innovate, with Kimi's OK Computer AI agent feature (released in September 2025), able to create full websites and slides from simple prompts.[334] More recently, Moonshot's new Kimi K2 Thinking model outperforms OpenAI's GPT-5 and Anthropic's Claude Sonnet 4.5 across multiple metrics.[335]
- **Baichuan**, founded in March 2023, shortly thereafter released two open-source, commercially usable Chinese language models: Baichuan-7B and Baichuan-13B.[336] These models have since been

329 Xiaomi not included as they are only active in FM development.
330 TechNode, 'Meet China's top six AI unicorns: who are leading the wave of AI in China', 9 January 2025.
331 Center for Data Innovation, 'Zhipu AI: China's Generative Trailblazer Grappling with Rising Competition', 12 December 2024.
332 Asia Tech Lens, 'Meet MiniMax: The Chinese Tech Company Touted by Jensen Huang That's Headed for an IPO', 30 July 2025. MiniMax, 'About MiniMax', accessed November 2025.
333 Center for Data Innovation, 'Moonshot AI: Betting Big on Long-Context, Confronting the Challenges of Scale and Reliability', 10 January 2025.
334 SCMP, 'Moonshot AI's Kimi assistant offers 'agent mode' for creating multi-page websites, slides', 25 September 2025.
335 SCMP, 'Why new model of China's Moonshot AI stirs 'DeepSeek moment' debate', 11 November 2025.
336 TechNode, 'Meet China's top six AI unicorns: who are leading the wave of AI in China', 9 January 2025.

updated and power Baichuan's chatbot Baixiaoying.[337] Baichuan also released its Baichuan-M1 (and later Baichuan M2 model) as the first open-source model designed for medical tasks.[338] This model is trained from scratch with a focus on medical data and is specifically optimised for medical scenarios.[339]

- **01.AI**, founded in July 2023, launched two models in 2025: Yi-Lightning and Yi-Large. These models have low training and deployment costs, with the Yi-Lightning model being trained on 2,000 GPUs and total training expenses of around 2% of xAI's training costs.[340]
- **StepFun**, founded in 2023, has launched 11 foundational model products, ranging in format from text and multimodal models, to video generation and speech models, including its open-source Step3 reasoning model.[341]

There are multiple other independent GenAI firms, most notably **DeepSeek** whose entry shocked the sector with innovations in model design and training processes. DeepSeek was founded in May 2023 by the quantitative hedge fund High-Flyer, rather than an established AI development laboratory.[342] DeepSeek's novel hardware-optimised algorithms and low training costs put it at the forefront of GenAI model development as discussed in section 2.4.1.2. Other Chinese firms that have developed their own models include Butterfly Effect, SenseTime, iFlytek, and ModelBest. Additionally, there are numerous specialised firms building their own proprietary LLMs using internal commercial data. These include Trip.com, China's largest travel platform that has built its proprietary LLM Wendao, and SF Technology that has built its Fengzhi model tailored to logistics operations.[343]

337 AsiaTechDaily, 'Alibaba-Backed Baichuan Secures $691 M, Valued at $2.7 Billion', 26 July 2024.
338 Baichuan AI, 'Baichuan-M2 Plus', accessed December 2025.
339 Baichuan AI, 'Baichuan-M1: Pushing the Medical Capability of Large Language Models', 18 February 2025.
340 TechNode, 'Meet China's top six AI unicorns: who are leading the wave of AI in China', 9 January 2025.
341 TechNode, 'Meet China's top six AI unicorns: who are leading the wave of AI in China', 9 January 2025.
Github, 'Step3', accessed December 2025.
342 Forbes, 'All About DeepSeek — The Chinese AI Startup Challenging US Big Tech', 26 January 2025.
343 Harvard Business Review, 'How Savvy Companies Are Using Chinese AI', October 2025.

Figure 11: Notable GenAI products from China's AI startups

Company	Notable GenAI models and apps
deepseek	• DeepSeek-V3 model • DeepSeek-R1 model • DeepSeek app
ZHIPU · AI	• GLM foundation models • ChatGLM chatbot app • AutoGLM AI assistant app
MINIMAX	• MiniMax M2 foundation model • Hailuo foundation models • MiniMax Agent app • Hailuo video app • Talkie app
Moonshot AI	• Kimi K2 foundation models • Kimi K2 Thinking model • Kimi chatbot app with OK Computer feature
百川智能 BAICHUAN AI	• Baichuan foundation models • Baichuan-M1 and Baichuan-M2 models • Baixiaoying chatbot app
01.AI	• Yi-Lightning foundation model • Yi-Large foundation model
StepFun	• Step3 foundation model • NextStep-1 foundation model • StepFun AI assistant app

Source: RBB summary based on desk research of publicly available sources.

5.1.2 DeepSeek's market disruption

The DeepSeek example shows that entry and expansion by a smaller, low-cost FM competitor is possible. The release of DeepSeek-V3 in December 2024 and DeepSeek-R1 in January 2025 disrupted the GenAI landscape. DeepSeek-R1 is considered to be as powerful as OpenAI's o1 model and was developed for a fraction of the cost.[344] DeepSeek used a combination of innovative techniques focused on computing efficiency to negate the effect of the lack of access to the most advanced AI chips. The technical components and performance details of DeepSeek's models are covered in section 2.4.1.2. As a result of these efficiency gains, DeepSeek offers prices that are significantly lower than those of its competitors.

DeepSeek's use of open-source releases with accompanying model weights and code repositories has influenced the commercial strategies of other market suppliers. For instance, Meta cited DeepSeek-R1 as validation of Meta's own open-source model strategy, noting that DeepSeek represented a key example of open-source models "*surpassing*" leading proprietary models.[345] Following the success of DeepSeek and the wider Chinese ecosystem's focus on open-source model development, OpenAI released GPT-OSS (Open Source Style) in August 2025, its first open-weight models since the open-source release of GPT-2 in 2019.[346]

344 Analytics Vidhya, 'DeepSeek R1 vs OpenAI o1: Which One is Faster, Cheaper and Smarter?', 4 April 2025.
345 Business Insider, 'Meta's chief AI scientist says DeepSeek's success shows that 'open source models are surpassing proprietary ones'', 26 January 2025.
346 CNBC, 'OpenAI's Altman warns the U.S. is underestimating China's next-gen AI threat', 18 August 2025.

DeepSeek's release of its chatbot app drove a surge in consumer demand. The DeepSeek app displaced OpenAI's ChatGPT as the most popular free iOS app in the US in January 2025.[347] DeepSeek's rapid surge in popularity suggests consumer use of GenAI is growing, and customers are willing to switch or multi-home, especially when quality improves or new features appear. This competitive threat contributed in part to the sell-off of AI stocks in US stock markets in late January 2025.[348] DeepSeek's low-cost approach poses a threat to competitors and creates pressure for efficiency-focused innovation in GenAI model development. DeepSeek's entrance affected both the domestic Chinese market and the international GenAI market putting competitive pressure on established firms to improve their prices and offerings, while providing a low-cost option for potential GenAI adopters facing resource constraints.

5.1.3 Price war

China's "war of a hundred models" led to aggressive price competition with both large tech firms and smaller, independent firms cutting prices to attract consumers and businesses.[349]

As summarised in Figure 12 below:

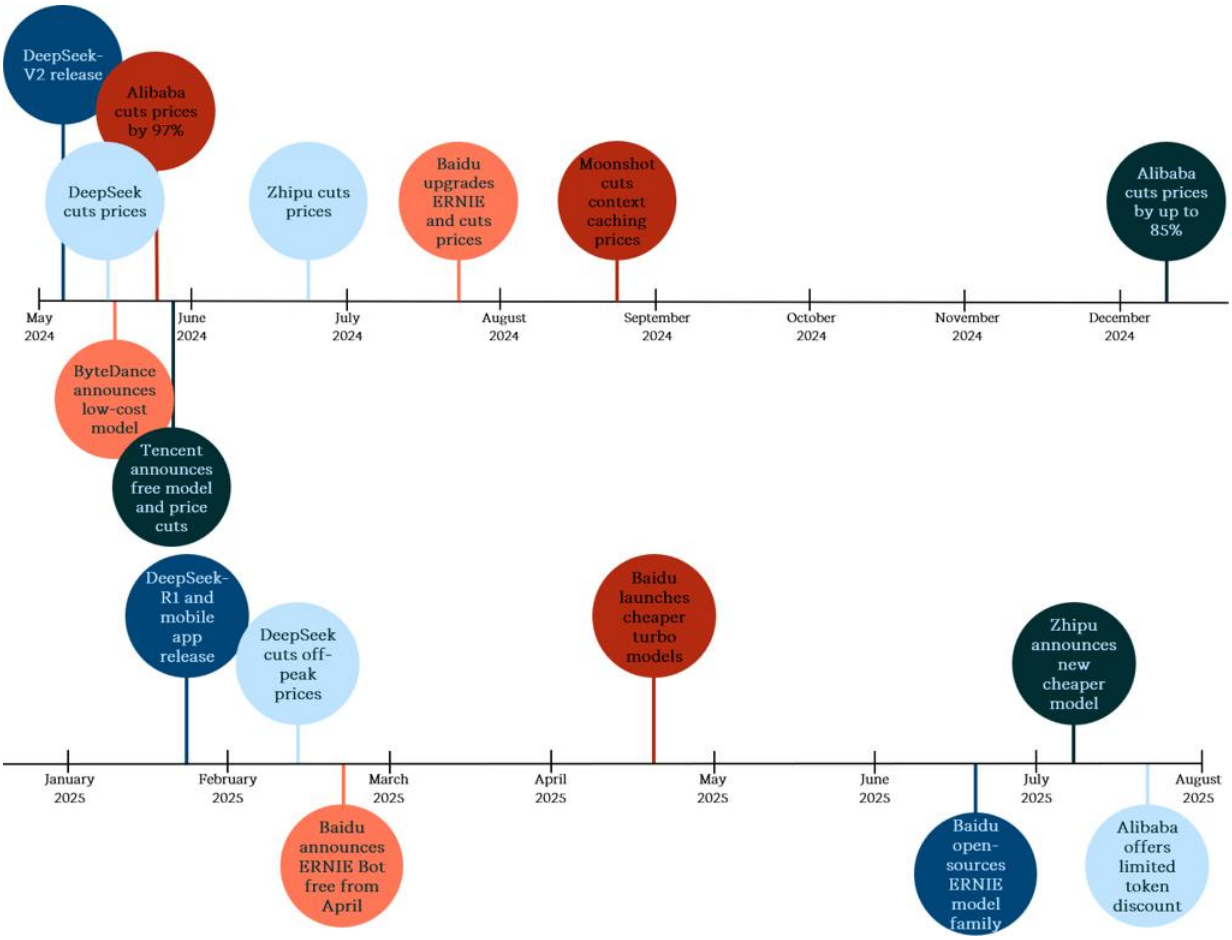
- In May 2024, DeepSeek released its chatbot model DeepSeek-V2 which quickly gained widespread popularity in China.[350] DeepSeek then triggered an industry-wide price war by cutting its model prices to roughly 1% of the price of GPT-4 Turbo, OpenAI's flagship model at the time.[351] In the same month, ByteDance introduced Doubao at a substantially lower price point, reinforcing the downward pressure on market pricing.[352] Alibaba reduced prices for its flagship model by up to 97%, while Tencent made the lite version of Hunyuan free and cut prices for higher-tier versions by between 50% and 88%.[353]
- In June 2024, Zhipu halved prices for its GLM series, resulting in prices around 90% below the prevailing industry level average.[354] Over the following months, Baidu upgraded its ERNIE model and implemented further price reductions and Moonshot halved context-caching costs (for storing and reusing data) for Kimi.[355] By the end of the year, Alibaba introduced another round of cuts (of up to 85%) across its LLM portfolio.[356]
- At the start of 2025, DeepSeek launched DeepSeek-R1 and its mobile chatbot app which brought the company to the forefront of international GenAI attention as discussed in section 5.1.2. In February, DeepSeek introduced discounted pricing for developers of up to 75% during off-peak hours; aligning with daytime hours in Europe and the United States.[357] Following this announcement, Baidu announced it would make its ERNIE chatbot free from April citing improvements in technology and reductions in costs.[358] Baidu later announced in April the launch of ERNIE 4.5 Turbo and ERNIE X1 Turbo, priced

[347] Engadget, 'China's DeepSeek AI assistant becomes top free iPhone app as US tech stocks take a hit', 28 January 2025.
[348] NVIDIA posted the largest one-day drop in market capitalisation on Wall Street. However, other tech firms such as Alphabet and Microsoft also saw stock price falls. This market dip was primarily driven by fears over the forecasts of future advanced semiconductor and cloud computing demand if infrastructure requirements for GenAI model training were to fall. However, some of this price drop can also likely be attributed to fears of DeepSeek providing competition to established players in the FM market.
Reuters, 'DeepSeek sparks AI stock selloff; Nvidia posts record market-cap loss', 28 January 2025.
Forbes, 'Nvidia Stock May Fall As DeepSeek's "Amazing" AI Model Disrupts OpenAI', 26 January 2025.
Britannica, 'DeepSeek', accessed November 2025.
[349] Reuters, 'China's AI 'war of a hundred models' heads for a shakeout', 22 September 2023.
[350] Britannica, 'DeepSeek', accessed November 2025.
[351] The Economist, 'A price war breaks out among China's AI-model builders', 13 June 2024.
[352] TechNode, 'ByteDance surprises AI rivals with ultra-low cost Doubao model', 16 May 2024.
[353] The Economist, 'A price war breaks out among China's AI-model builders', 13 June 2024.
Reuters, 'Tencent and iFlytek enter China's AI language model price war', 22 May 2024.
[354] SCMP, 'Tech unicorn Zhipu AI joins China's LLM price war amid new funding round', 5 June 2024.
Center for Data Innovation, 'Zhipu AI: China's Generative Trailblazer Grappling with Rising Competition', 12 December 2024.
[355] Bloomberg, 'Baidu Upgrades Ernie AI Model, Cuts Pricing Further', 5 July 2024.
SCMP, 'Chinese AI start-up Moonshot cuts LLM feature price amid fierce domestic competition', 8 August 2024.
[356] SiliconANGLE, 'Alibaba Cloud announces aggressive LLM price cuts in bid to dominate China's AI market', 1 January 2025.
[357] Reuters, 'DeepSeek cuts off-peak pricing for developers by up to 75%', 26 February 2025.
[358] Reuters, 'Baidu to make AI chatbot Ernie Bot free of charge from April 1', 13 February 2025.

at 20% and 50% of their non-turbo counterparts respectively, and provided its ERNIE 4.5 model family as open-source in June 2025.[359]

- In July 2025, Zhipu introduced its new open-source model, GLM-4.5, positioning it as a lower-cost alternative to DeepSeek. Specifically, Zhipu priced GLM-4.5 input and output tokens approximately 20% and 90% lower than those of DeepSeek-R1, respectively.[360] In the same month, Alibaba announced a limited-time 75% discount on cached input tokens (for previous prompts stored for reuse) for its Qwen3-Coder-Plus model.[361]

Figure 12: Timeline of key price cuts in China's "war of a hundred models"



Source: RBB summary based on desk research of publicly available sources.

Amidst escalating price competition, cross-platform integration among Chinese GenAI firms has continued to increase, with more firms incorporating competitors' models into their own apps and search products, broadening consumer and developer access to the latest models. For example, DeepSeek has been integrated into several competitors' platforms: Tencent has beta-tested DeepSeek's AI-search capabilities within its popular app Weixin, and Baidu has integrated DeepSeek alongside its ERNIE model in Baidu Search.[362] Major technology platforms, such as Alibaba, Tencent, Baidu, and ByteDance, all offer multiple competing models through their cloud platforms despite developing and promoting their own models (see section 4.1.1).

359 PR Newswire, 'Baidu Launches ERNIE 4.5 Turbo, ERNIE X1 Turbo and New Suite of AI Tools to Empower Developers and Supercharge AI Innovation', 25 April 2025.
ERNIE, 'Announcing the Open Source Release of the ERNIE 4.5 Model Family', 30 June 2025.

360 Zhipu announced that GLM-4.5 would be priced at USD 0.11 per million input tokens and USD 0.28 per million output tokens. DeepSeek-R1, by comparison, was priced at USD 0.14 per million input tokens and USD 2.19 per million output tokens.
CNBC, 'China's latest AI model claims to be even cheaper to use than DeepSeek', 28 July 2025.

361 Alibaba Cloud, 'Limited-time discount for Qwen3-Coder-Plus', 8 August 2025.

362 Reuters, 'Tencent's Weixin app, Baidu launch DeepSeek search testing', 17 February 2025.
TechNode, 'Baidu Search integrates DeepSeek and Large Model ERNIE for advanced search', 17 February 2025.

5.1.4 Implications for FM competition

China's FM sector demonstrates four distinctive competitive patterns.

Severe price competition and performance parity. The 2024-2025 price war saw dramatic API cost reductions, with multiple providers offering entirely free access: Baidu ERNIE Speed and ERNIE Lite (free for enterprise users), ByteDance Doubao Pro (USD 0.0001 per 1,000 tokens), Alibaba Qwen models (up to 97% price cuts), and Tencent Hunyuan-lite (free with discounted access for the standard model).[363] Industry analysts attributed this convergence to technological parity: "*LLM technologies and capabilities have already converged so there are no particularly strong barriers*," with price competition predicted to "squeeze profit margins."[364]

Regulatory batch approval intensified competitive dynamics by creating simultaneous rather than sequential market entry. The August 2023 approval wave granted multiple major competitors (including Baidu, ByteDance, Baichuan Intelligence, SenseTime and Zhipu AI) simultaneous commercial deployment authorisation, forcing immediate head-to-head competition rather than the sequential launch patterns observed in other markets where early entrants establish user bases before later competitors receive approval.[365] This regulatory procedural choice created flash competition where multiple well-resourced providers launched consumer-facing services simultaneously, accelerating competitive intensity compared to markets where temporal sequencing could enable potential first-mover advantages.

Sustained competitive plurality. Over 200 organisations attempted LLM launches, with 117 receiving regulatory approval by March 2024 and 238 GenAI services being approved in 2024 alone.[366] By April 2025, China saw a total of 346 GenAI services registered with 5-7 major competitors (Baidu, Alibaba, Tencent, ByteDance, DeepSeek, Moonshot and Zhipu) capturing majority usage while dozens of smaller providers maintained a market presence serving specialised segments.[367] This pattern demonstrates that China, to date, can sustain multiple competitors.

Chinese models achieved global competitive recognition by 2025, showing that domestic competition led to frontier innovation. By mid-August 2025, Alibaba's Qwen3-Coder accounted for over 20% of model usage on OpenRouter, second only to Anthropic's Claude Sonnet 4 at over 30%.[368] Qwen models led Hugging Face in open LLM leaderboards across multiple categories.[369] Chinese models topped UC Berkeley's Chatbot Arena benchmarks ahead of offerings from Google and Meta.[370] DeepSeek's V3 and R1 models demonstrated frontier capabilities, with R1 displacing ChatGPT as the

---

363 TechNode, 'Price war in China's AI field expanded as Alibaba Cloud and Baidu have drastically reduced model prices', 22 May 2024.
Fortune, 'The race to lead China's AI sector heats up as ByteDance, Alibaba and Baidu offer their models at rock-bottom prices', 22 May 2024.
Yicai Global, 'Tencent, iFlytek Join China's LLM Price War', 23 May 2024.
PYMNTS, 'China's AI Price War May Spark Global Tech Showdown, Industry Insiders Say', 23 August 2024.
SCMP, 'TikTok owner ByteDance launches low-cost Doubao AI models for enterprises, initiating a price war in crowded mainland market', 15 May 2024.

364 Yicai Global, 'Tencent, iFlytek Join China's LLM Price War', 23 May 2024.

365 Bloomberg, 'Baidu Among First Firms to Win China Approval for AI Models', 30 August 2023.
SCMP, 'Baidu, SenseTime open AI chatbots to the public after China grants first approvals for such services', 31 August 2023.
Rappler, 'China lets Baidu, others launch ChatGPT-like bots to public, tech shares jump', 31 August 2023.
Sixth Tone, 'China's Top AI Trends in 2023', 26 December 2023.

366 These GenAI services are public-facing applications which fall under the scope of Article 2 of the Interim Measures for the Administration of Generative Artificial Intelligence Services. Consequently, downstream chatbots of Chinese FMs are subject to this regulation.
Tom's Hardware, 'China's AI model glut is a 'significant waste of resources' due to scarce real-world applications for 100+ LLMs says Baidu CEO', 8 July 2024.
China News, 'China's generative AI service users surpass 600 million: report', 1 August 2025.

367 TechNode, 'China approves 346 generative AI services under national registration scheme', 9 April 2025.
Cyberspace Administration of China, 'Interim Measures for the Administration of Generative Artificial Intelligence Services', 13 July 2023.

368 Dataconomy, 'Alibaba Qwen 3 Coder Gains 20% Share On OpenRouter', 19 August 2025.
OfficeChai, 'Alibaba's Qwen3 Coder Appears To Take Market Share From Anthropic & Google As Per OpenRouter Data', 18 August 2025.

369 Hugging Face, 'Model statistics of the 50 most downloaded entities on Hugging Face', 13 October 2025.

370 LM Arena, 'LMSYS Chatbot Arena leaderboard', accessed on various dates 2024-2025.

number one free iOS app in the United States in January 2025.[371] Meta's Chief AI Scientist at the time, Yann LeCun, stated that DeepSeek-R1's performance demonstrated that "*open source models are surpassing proprietary ones*".[372]

However, regulatory concerns about data security have led to restrictions in some markets: Australia and Taiwan banned DeepSeek from government devices, while Italy imposed an outright ban on the app, echoing restrictions previously applied to ByteDance's TikTok.[373] These regulatory barriers potentially limit international market access despite technical competitiveness.

This international market penetration, albeit constrained by regulatory restrictions in some jurisdictions, indicates China's competitive dynamics produced models achieving technical parity with frontier models while operating under export controls limiting access to advanced chips, suggesting competitive intensity can drive efficiency innovations in response to resource constraints.

5.2 GenAI deployment case studies in the Asia-Pacific region

5.2.1 OpenAI's China exit

On 24 June 2024, OpenAI announced additional restrictions to block API traffic from regions where access to OpenAI services was not supported.[374] This announcement meant that developers in China would lose access to all OpenAI platforms on 9 July 2024.[375]

While OpenAI's consumer-facing product, ChatGPT, was not available in China even before that date, several Chinese developers relied on API to access OpenAI products to build their own applications.[376] The block primarily affected two categories of Chinese developers: those building applications powered by OpenAI's models and local AI companies using OpenAI outputs to train their own models.[377]

As a result, in the short two-week window between the announcement and the discontinuation of the service, Chinese developers had to switch to alternatives to prevent disruption to their applications. This event represented a unique natural experiment with respect to GenAI developers' abilities to switch FM providers within a limited timeframe.

As covered in the competitive implications (section 4.1.2), this natural experiment of OpenAI's exit provides evidence that switching FM providers is feasible and the costs are relatively limited as many viable competitors offer similar alternatives.

5.2.1.1 Scrambling to fill the void

Immediately following the announcement, several local competitors used this opportunity to attract outgoing OpenAI customers to their own platforms.[378]

---

371 Reuters, 'DeepSeek sparks AI stock selloff; Nvidia posts record market-cap loss', 28 January 2025.
Financial Times, 'Tech stocks slump as China's DeepSeek stokes fears over AI spending', 28 January 2025.
372 Business Insider, 'Meta's chief AI scientist says DeepSeek's success shows that open source models are surpassing proprietary ones', 25 January 2025.
373 Britannica Money, 'DeepSeek', accessed November 2025.
374 Reuters, 'Chinese AI firms woo OpenAI users as US company plans API restrictions', 26 June 2024.
375 Rest of World, 'OpenAI cuts its last and most important link to China', 27 June 2024.
376 Ibid.
377 ChinaTalk, 'OpenAI Pulls the Plug on China', 11 July 2024.
378 Rest of World, 'OpenAI cuts its last and most important link to China', 27 June 2024.
Reuters, 'Chinese AI firms woo OpenAI users as US company plans API restrictions', 26 June 2024.

Multiple firms offered switching incentives, promising to match the scale of the developers' existing OpenAI usage, or offering free or heavily discounted tokens.[379] In total, more than ten providers reacted to the announcement with migration offers.[380]

- Baidu said it would launch an "inclusive program" offering new users 50 million free tokens for its flagship ERNIE 3.5 model. In the announcement, Baidu specifically targeted existing OpenAI users, offering a token package equivalent to their scale of OpenAI usage.[381]
- Alibaba Cloud also announced free tokens and migration through its AI platform.[382] It emphasised having a more cost-effective LLM alternative to ChatGPT for users in China and offered 22 million free tokens plus exclusive migration services for app developers.[383]
- Zhipu AI, another major player in China's AI sector, announced a "Special Migration Program" for OpenAI API users.[384] In the announcement, Zhipu emphasised that its GLM model performs on a similar level to OpenAI products.[385] The announcement boasted the largest usage offer among Chinese developers, with up to 150 million free tokens and a series of training sessions.[386]
- Despite a global partnership with OpenAI, Microsoft also published a step-by-step guide to how to migrate to its local service, operated by local partner 21Vianet.[387] Under the proposed solution, developers from China could continue using OpenAI services even after the shutdown date by using Microsoft's Azure.[388]

Several of these providers offered "one-click migration" services that ensured a smooth "relocation" processes for the affected firms. As early as 2 July 2025, it was reported that the impact of the ban was minimal.[389] Instead, the OpenAI exit offered local players an opportunity to attract former OpenAI users and grow their user base, rather than resulting in widespread disruption in Chinese GenAI applications as would have occurred if developers had struggled to migrate.[390]

5.2.1.2 Further restrictions

OpenAI is not the only GenAI firm that restricted the use of its products in China. Another leading US developer, Anthropic, has taken extensive measures against access to its AI products from certain restricted regions, including China. In September 2025, Anthropic prohibited access to its services for organisations whose ownership structures indicate ultimate control by entities headquartered in jurisdictions where Anthropic's products are not permitted (such as China), even if those organisations operate outside those regions.[391] This led to a further shift away from US models in China by some companies in the wider APAC region, with a notable example being the Singapore-based coding app Trae (owned by Chinese ByteDance), which removed Anthropic's Claude models from its products.[392]

---

379 Rest of World, 'OpenAI cuts its last and most important link to China', 27 June 2024.
Reuters, 'Chinese AI firms woo OpenAI users as US company plans API restrictions', 26 June 2024.
Time, 'How OpenAI's Decision Not to Operate in China Will Reshape the Chinese AI Scene', 26 June 2024.
380 This includes Baidu, Alibaba, Tencent, Zhipu AI, 01.AI, Baichuan, SenseTime Group, 21Vianet, iFlytek, Moonshot AI, Minimax.
381 Baidu, 'Hometown Clouds: Domestic Large Model Universal Benefit Plan', 25 June 2024.
Reuters, 'Chinese AI firms woo OpenAI users as US company plans API restrictions', 26 June 2024.
382 Reuters, 'Chinese AI firms woo OpenAI users as US company plans API restrictions', 26 June 2024.
383 Yicai Global, 'Alibaba, Baidu, Other Chinese AI Firms Rush to Fill Gap Left by ChatGPT in China', 26 June 2024.
384 Reuters, 'Chinese AI firms woo OpenAI users as US company plans API restrictions', 26 June 2024.
385 Ibid.
386 Time, 'How OpenAI's Decision Not to Operate in China Will Reshape the Chinese AI Scene', 26 June 2024.
387 Ibid.
388 Yahoo Tech, 'Microsoft Azure China offers Chinese businesses a loophole to OpenAI's departure', 9 July 2024.
389 STCN, '"Relocation" or "Going Global": OpenAI's Countdown to the Discontinuation of its Chinese APIs', 2 July 2024.
390 Time, 'How OpenAI's Decision Not to Operate in China Will Reshape the Chinese AI Scene', 26 June 2024.
391 Anthropic, 'Updating restrictions of sales to unsupported regions', 5 September 2025.
392 SCMP, 'Tech war: ByteDance cuts off Claude model after Anthropic restricts China access', 5 November 2025.
SCMP, 'Impact of Anthropic's China ban unlikely to 'move the needle', expert says', 7 October 2025.

These developments led to a concern that some open-source models, like Meta's Llama, would also be cut-off following political tensions between the US and China.[393] However, unlike proprietary APIs that can be restricted through terms of service, the open-source nature of Llama models makes access controls difficult to enforce. Chinese developers continue using and adapting open-source models, with localised versions such as Chinese-Llama-Alpaca providing more Chinese vocabulary and incorporating local training data.[394]

5.2.2 Grab's integration of AI into its super app

Grab is a super app active in eight Southeast-Asian countries, which combines a variety of services in a single platform, including ride-hailing, food delivery and grocery delivery. It also provides financial services like GrabPay, GrabFinance and GrabInsure.[395] Grab has 46 million monthly active users and is present in over 500 cities.[396]

Grab uses AI tools internally to improve business and product development initiatives and develops user-facing features based on both home-grown machine learning and third-party GenAI tools. Grab's home-grown AI tools include the KartaCam2 camera (the newest generation of Grab's cameras used for mapping), Mystique AI (Grab's internal copywriting AI) and Vision LLM (developed to accurately read and process Southeast Asian scripts).

As covered in the competitive implications (sections 4.1.1, 4.1.3 and 4.5.2), Grab's observed multi-homing across FM providers illustrates that platform competition between super apps encourages strategic model deployment based on use case in order to provide competitive AI features. Additionally, Grab's development of Vision LLM shows that the linguistic diversity of the APAC regions can result in competitive advantages for local model developers.

5.2.2.1 Third-party AI tools in the Grab app

AI Merchant Assistant

The AI Merchant Assistant is a chatbot incorporated in the GrabMerchant app that allows suppliers to ask business-related questions.

Grab uses OpenAI's LLM to perform analyses and provide tailored recommendations for merchants, while the user interface, which integrates chat functionalities, is powered by Anthropic's Claude.[397] This allows the suppliers to receive personalised responses and advice on their business delivered in a human tone. It can also contact suppliers proactively and offer advice or help with updating menus and creating ad campaigns. Since shifting to AI-based merchant support, the tool has experienced an increase in inbound merchant conversations by 8% (and a 5.7 percentage point increase in the resolution rate).[398]

This integration of AI features extends to Grab's Merchant Menu Assistant which allows new businesses to upload a menu to the Grab app by taking a picture of their physical menu.[399] This feature was developed internally at a Grab Hackathon.[400] This feature was further extended and suppliers can now generate AI dish descriptions, improving checkout rates.[401] Grab plans additional GenAI functionalities within the AI

393 Time, 'How OpenAI's Decision Not to Operate in China Will Reshape the Chinese AI Scene', 26 June 2024.
394 Source Forge, 'Chinese-LLaMA-Alpaca-2 v2.0', accessed November 2025.
395 Grab, 'Grab Financial Group', accessed November 2025.
396 This figure is based on monthly transacting users. Grab, 'Grab Reports Second Quarter 2025 results', 31 July 2025. Grab, 'Where We Are', accessed November 2025.
397 Grab, 'Grab deploys agentic AI to empower merchants and driver partners', 8 April 2025.
398 Claude, 'Grab scales personalized merchant support across Southeast Asia with Claude', accessed November 2025.
399 Grab, 'Grab's new AI tool makes menu management a snap for merchants', 26 November 2024.
400 Ibid.
401 Phocus Wire, 'Grab revenue up 17% YoY, focused on AI future growth', 15 August 2024.

Merchant Assistant in the near future, such as recommending customised financing solutions for merchants, including products from GrabFin and Grab's digital banks.[402]

AI Driver Companion

The AI driver Companion tool is incorporated in the existing Grab Driver app and is underpinned by OpenAI FMs. It incorporates the following features.

- **Road conditions:** This tool allows drivers to flag road issues such as traffic or roadworks verbally, ensuring driver safety. It uses OpenAI's GPT-4o to process the driver reports accurately and OpenAI's Real Time API technology to respond in a natural tone. Since the pilot rollout, the tool has received over 16,000 reports a day from over 900,000 drivers across the region.[403]
- **Ride guidance:** This tool analyses current traffic, user data, historical demand, and the weather (among other factors), to forecast ride demand by area, and guide drivers towards likely hotspots. Since its implementation, over 250,000 drivers have used the feature every week.[404] It improves the efficiency of the platform by reducing the wait-time for drivers and riders, increasing driver earnings and user satisfaction.
- **AI consumer notes:** Integrated into GrabMaps, this feature extracts specific instructions from custom-written delivery notes and combines these with images from crowdsourced map data to provide clear, real-time instructions to the driver.[405]

AI user interface

Grab and OpenAI are exploring the applications of AI to customer support and user interactions, in an effort to make the app accessible to a wider range of users. The voice capabilities of OpenAI will allow Grab to improve their customer experience and simplify access for the visually impaired and elderly who may otherwise find it difficult to navigate the on-screen app interface.[406]

Grab is also exploring GenAI customer support chatbots that can solve users' issues swiftly, by introducing Standard Operating Procedures (**SOP**)-driven LLM agent frameworks.[407] This uses LLMs to respond to complex customer issues, but provides a step-by-step framework to guide LLM responses. This is especially helpful in cases where users contact customer service to flag fraudulent behaviour or transactions. The core idea is to ground GenAI outputs by ensuring adherence to pre-defined SOPs, breaking the inference process into smaller, simplified steps and reducing the risk of hallucinations.[408]

5.2.2.2 Vision LLM

Grab's Vision LLM processes documents like IDs and driver licenses to verify customer, merchant, or rider identities. Existing AI and Western LLMs have difficulties reading and processing Southeast Asian scripts like Bahasa Indonesia, Thai and Vietnamese. In addition, traditional Optical Character Recognition (**OCR**) systems are error-prone in this context due to the variety of documents and scripts to be scanned.

To combat this, Grab developed its own vision LLM which accurately "reads" Southeast Asian scripts and can extract the necessary information accurately.[409] To optimise resources and create a model tailored

402 Grab, 'Grab deploys agentic AI to empower merchants and driver partners', 8 April 2025.
403 Ibid.
404 Ibid.
405 Grab, 'How Grab uses AI to generate more precise delivery instructions', 29 January 2024.
406 Grab Investor Relations, 'Grab and OpenAI announce strategic collaboration, first of its kind in Southeast Asia', 30 May 2024.
407 Grab Tech Blog, 'Introducing the SOP-driven LLM agent frameworks', 25 April 2025.
408 Ibid.
409 A Vision LLM is an AI tool which can be fed images as input and can transform this into information which can be processed by the LLM together with any other text input.

to their needs, Grab built a lightweight Vision LLM (with around 1 billion parameters).[410] Grab tested several open-source models and chose Qwen2-VL 2B multimodal LLM because of its efficient size (allowing full fine-tuning with limited GPU memory), effective Southeast Asia (**SEA**) language tokeniser (for Thai and Vietnamese), and support for dynamic image resolution (essential for high-accuracy OCR).[411, 412]

While Grab used an existing open-source LLM as the foundation, the process involved multiple fine-tuning stages to improve Southeast Asian language script recognition and achieve the desired standard of accuracy. Additionally, the Grab developer team noted superior latency compared to other OCR models such as Qwen2, ChatGPT and Google's Gemini.[413]

5.2.2.3 Productivity tools

Grab also uses a variety of AI tools for internal purposes across different areas. These includes internal GenAI tools based on OpenAI's models, which were trained with internal information, as well as off-the-shelf GenAI tools that offer specific functionalities to Grab's employees.

To enhance efficiency, the Grab Data Analytics team developed Mystique AI, a tool used for copywriting.[414] It uses a mix of technologies including machine learning, LLMs and RAG to mimic the style of Grab copywriting and tailor messages to the specific user.[415] For example, it can formulate targeted messages to regular users suggesting their standard order, or craft a message to an inactive user to encourage engagement. Mystique AI can provide different tones based on the context to improve customer engagement.

Grab's design team developed Mosaic, which generates illustrations in Grab's style. It was initially trained using around 40,000 training steps and continues to be improved.[416] This tool increases the design team's productivity considerably, reducing a day of an illustrator's work to just 15 seconds, and is currently used to produce about 800 illustrations daily.[417]

Grab's sales team is also piloting the use of Einstein AI (provided by Salesforce). Grab is piloting how to use Einstein AI to get certain insights out of the Salesforce database to help salespeople improve their interactions with customers.[418]

In engineering, Grab uses a variety of AI-powered coding assistants for tasks ranging from code autocompletion and error detection to the generation of data to test new code. Grab specifically decided to multi-home and use multiple tools to have different options for different problems, and to be able to change tool quickly as innovation leads to one tool being better than another for a particular task.[419] Despite these tools being a recent phenomenon, over half of Grab's data engineers use AI assistants regularly. Different tools provide more value in areas where they perform better.

- **ChatGPT Enterprise:** used by Grab's software engineers as part of a multi-tool approach to enhance their engineering workflow.

---

410 Grab Tech Blog, 'How we built a custom vision LLM to improve document processing at Grab', 4 November 2025.
411 Note, the Grab developer team initially experimented with fine-tuning a base Qwen model and compressing this model using Low - Rank Adaptation techniques (LoRA). Ultimately the team achieved better performance from using different base models for vision encoding (Qwen2-VL 2B), language decoding (Qwen2.5 0.5B), and the projection layer (customised solution).
Grab Tech Blog, 'How we built a custom vision LLM to improve document processing at Grab', 4 November 2025.
412 The Register, 'LLMs are lousy at reading Asian languages, finds Singapore's Grab', 4 November 2025.
413 Ibid.
414 Grab, 'Meet Mystique, the AI-powered copywriting tool that helps Grab communicate effectively', 6 September 2024.
415 Grab, 'The AI technology that gives Grab's copywriting tool Mystique its voice', 6 September 2024.
416 Grab, 'How the design team at Grab is redefining creativity with AI', 12 June 2025.
417 Ibid.
418 Salesforce, 'How Grab is Using Einstein AI to Secure First-Mover Advantage in Southeast Asia', accessed November 2025.
419 Grab, 'Beyond one-size-fits-all: Why Grab embraces multiple AI coding assistants', 20 February 2025.

- **GitHub Copilot:** used by engineers for the quick autocompletion of codes and to generate mock data for testing code.
- **Sourcegraph Cody:** helps engineers understand large codebases and track function changes across services.
- **Cursor:** has code completion capabilities and can make targeted improvements that improve code readability.

Grab acknowledges that the long-term impact of its AI coding assistants is still unclear and remains open to revisiting its approach. The decision to employ multiple tools was driven by the flexibility this approach offers and the possibility to evaluate the impact different tools have in real-time.[420]

Grab also recently launched an AI centre of excellence in Singapore. The centre is backed by Digital Industry Singapore and has partnered with the Singapore Association of the Visually Handicapped and Singapore's national water authority.[421] The centre has developed a voice-assistant feature which allows the visually impaired or elderly to use Grab to book rides. While this tool is currently based on OpenAI models, one of the key projects for Grab is to build its own GenAI FM, trained on its own data, which will result in better suggestions for routes, drivers and users.[422]

5.2.3 Samsung strategy across consumer and enterprise markets

Samsung Electronics pursues three distinct approaches to FM deployment:

- the company builds its own Gauss models for internal operations;
- partners with multiple external providers (Google Cloud's Gemini for consumer devices, Baidu's ERNIE Bot in China, and Microsoft's Azure AI for customer service); and
- through its subsidiary Samsung SDS, distributes multiple third-party models via the FabriX enterprise platform.[423]

This deployment spans different use cases across the three approaches. Gauss handles internal productivity and code development, Gemini powers consumer-facing Galaxy AI features, while FabriX offers enterprise customers access to ChatGPT, Llama, HyperClova X, and Samsung's own models.[424] The three approaches work together, showing how vertically integrated technology companies balance building their own technology with partnering with others to give customers a choice.

5.2.3.1 Three-pronged approach to FM deployment

Samsung's integrated ecosystem generates proprietary data across consumer devices (smartphones, tablets and wearables serving hundreds of millions of users), enterprise services (through FabriX customer deployments), and internal operations spanning over 70 countries.[425] This cross-domain data,

---

420 Ibid.
421 Vulcan Post, 'Grab launches its first AI Centre of Excellence, plans to hire at least 50 "high-value" roles', 23 May 2025.
422 Techwire Asia, 'Grab opens AI Centre in Singapore to tackle real-world challenges', 26 May 2025.
Grab, 'Grab Launches First Artificial Intelligence Centre of Excellence with Support from Digital Industry Singapore', 23 May 2025.
423 Samsung Research, 'Artificial Intelligence', accessed November 2025.
Samsung SDS, 'Samsung SDS Unveils Generative AI Services "FabriX" and "Brity Copilot" to Drive Hyperautomation in Corporate Business', 8 May 2024.
Samsung, 'Samsung and Google Cloud Join Forces To Bring Generative AI to Samsung Galaxy S24 Series', 18 January 2024.
Techwire Asia, 'Baidu AI Cloud partners Samsung Electronics China on generative AI smartphone', 29 January 2024.
Bloomberg, 'Samsung to Showcase Baidu's Ernie AI in Latest Galaxy Phones', 29 January 2024.
424 Samsung Research, 'Artificial Intelligence', accessed November 2025.
Korea Economic Daily, 'Samsung SDS AI service FabriX to launch on Microsoft Azure platform', 3 September 2024.
Samsung SDS, 'Fabrix', accessed December 2025.
The Investor, 'IT solutions provider aims to heighten work efficiency by innovating hyperautomation', 13 September 2023.
Samsung SDS, 'Samsung SDS Realizes Hyper-Automation Innovation with GPU-Centric AI Cloud', 4 September 2024.
425 Samsung, 'Fast-Facts', accessed December 2025.

combined with the in-house expertise in hardware and software integration, enables Samsung to train specialised models for specific use cases and integrate third-party solutions into its ecosystem.

Build: Internal development of Samsung Gauss

Samsung introduced Samsung Gauss in November 2023.[426] The model family includes three specialised variants: Samsung Gauss Language for text generation and comprehension, Samsung Gauss Code for software development assistance, and Samsung Gauss Image for visual content creation and manipulation.[427]

In November 2024, Samsung released Gauss 2, which, according to the company, performs 1.5 to 3 times faster than the original version while supporting 9 to 14 languages and various coding languages.[428] Gauss 2 deploys in three configurations: Compact (a small-sized model designed for on-device use), Balanced, and Supreme (using the Mixture of Experts (**MoE**) technology, which divides an AI model into multiple "experts" that specialise in a subset of the input data).[429]

Samsung has deployed Gauss 2 extensively for internal operations, with over 60% of developers in Samsung's Device eXperience division using the model through "code.i," an internal coding assistant.[430] The model is also deployed in call centres where it categorises and summarises customer interactions, while the Samsung Gauss Portal provides conversational AI to employees across the Device eXperience division for tasks including email composition, document summary, and translation.[431]

Buy: External partnerships with multiple FM providers

Samsung deploys FMs from three major external providers across different use cases and regions. Each partnership reflects specific optimisation criteria based on technical capabilities, regional regulatory requirements, and use-case specialisation.

In January 2024, Samsung and Google Cloud announced a partnership integrating Gemini into the Galaxy S24 series as the foundation for Galaxy AI features.[432] Samsung selected Gemini following "months of rigorous testing and competitive evaluation", which suggests that the company considered alternative providers.[433] Samsung was also among the first customers selected to test Gemini Ultra, Google's most capable model, though the S24 launch deployed Gemini Pro and Nano (a smaller on-device LLM).[434] Gemini powers consumer-facing features including text summaries across Samsung Notes, Voice Recorder, and Keyboard, along with image generation capabilities through Imagen 2 for photo editing.[435] Samsung planned to deploy Galaxy AI across 200 million devices by the end of 2024, expanding to 400 million devices by the end of 2025.[436] By mid-2025, Samsung reported that over 70% of Galaxy S25 users engaged with Galaxy AI features.[437]

The partnership operates differently in different markets. In China, where Google services are blocked, Samsung partnered with Baidu instead, using Baidu's ERNIE Bot to power GenAI features in Chinese

---

426 Samsung, 'Samsung Developer November Newsletter', 28 November 2023.
427 Ibid.
428 Samsung, 'Samsung Electronics Hosts Samsung Developer Conference Korea 2024, Unveils Its Improved Gen AI Model', 21 November 2024.
429 Ibid.
430 Ibid.
431 Ibid.
432 Samsung, 'Samsung and Google Cloud Join Forces To Bring Generative AI to Samsung Galaxy S24 Series', 18 January 2024.
433 Ibid.
434 Ibid.
435 Ibid.
436 Samsung Mobile Press, 'Samsung Continues to Expand Galaxy AI as Consumers Show Increasing Reliance on Mobile AI', 11 July 2025.
437 Ibid.

Galaxy devices.[438] The partnership, announced in January 2024, provides the same functional capabilities as Gemini in other markets: text summaries, real-time call translation, and Circle to Search functionality.[439] This regulatory-driven substitution demonstrates that deployers must maintain relationships with multiple providers to ensure regional compliance.

In September 2024, Samsung deployed Microsoft's GenAI technology for customer-facing service applications.[440] Following internal technical evaluation of Microsoft, Google and Palantir solutions during July 2024, Samsung selected Microsoft Azure AI Studio to build AI-powered chatbot systems for Samsung.com and retail stores in South Korea and the UK.[441] These chatbots provide product function explanations and purchase guidance. Samsung's evaluation process involved month-long proof-of-concept testing where employees used all three suppliers' solutions internally, with Microsoft achieving the highest performance scores.[442] This partnership demonstrates that Samsung continues evaluating and adding providers even after establishing major partnerships, and that Samsung selects different providers for different customer interactions based on use-case-specific performance criteria.

Samsung's TM Roh has summarised Samsung's approach: "*We are working with various partners for our on-device and cloud-based AI", "[w]e have our own [large language model] as well"* and *"we can put the optimization in there and we can also collaborate with our partners*."[443] This multi-homing approach operates simultaneously across consumer devices (Google globally, Baidu in China), customer service (Microsoft) and internal operations (Gauss).

Platform: **Samsung SDS FabriX multi-model enterprise distribution**

Samsung SDS launched FabriX in May 2024 as an enterprise GenAI service platform.[444] Samsung SDS' Scott HJ Koo described FabriX as "*Korea's largest generative AI service platform for enterprise use*" and has indicated that it was used by nearly 100,000 people.[445] By September 2024, approximately 100 corporate customers had adopted FabriX and Brity Copilot (Samsung's suite of corporate AI-powered productivity tools), with over 150,000 users.[446]

FabriX connects GenAI capabilities with enterprises' internal systems.[447] This allows enterprises to deploy AI across their operations without building custom infrastructure for each separate model integration. Notably, FabriX can be linked with various LLMs, including OpenAI models such as ChatGPT and HyperClova X from Naver, a competitor in South Korean AI markets.[448] This approach provides enterprise customers with choice in model selection for different use cases.

Enterprise clients using FabriX span multiple sectors, including Paradise Group and Woongjin for business efficiency improvements.[449] Other clients include Vietnam's CMC Group, Korean Air, LIG Nex1, and South Korea's Rural Development Administration.[450] Samsung SDS also offers Brity Copilot alongside FabriX, which automates business processes including email, messaging, meetings, and document management. Samsung SDS reports using Brity Copilot internally, noting time savings for office work such as composing

---

438 Techwire Asia, 'Baidu AI Cloud partners Samsung Electronics China on generative AI smartphone', 29 January 2024.
Bloomberg, 'Samsung to Showcase Baidu's Ernie AI in Latest Galaxy Phones', 29 January 2024.
439 Ibid.
440 Electronic Times, 'Samsung Electronics selects Microsoft as AI supplier for customer service', 2 September 2024.
441 Ibid.
442 Ibid.
443 AI Buzz, 'Samsung's Multi-Provider AI: Moving Galaxy AI Beyond Gemini', 25 July 2025.
444 Samsung SDS, 'Samsung SDS Unveils Generative AI Services "FabriX" and "Brity Copilot" to Drive Hyperautomation in Corporate Business', 8 May 2024.
445 Ibid.
446 Samsung SDS, 'Samsung SDS Realizes Hyper-Automation Innovation with GPU-Centric AI Cloud', 4 September 2024.
447 Samsung SDS, 'Samsung SDS Unveils Generative AI Services "FabriX" and "Brity Copilot" to Drive Hyperautomation in Corporate Business', 8 May 2024.
448 Samsung SDS, 'Samsung SDS to Lead Innovation in Hyperautomation with Generative AI', 14 September 2023.
449 The Korea Economic Daily, 'Samsung SDS AI service FabriX to launch on Microsoft Azure platform', 3 September 2024.
450 Ibid.

and summarising emails. Internal testing suggested that Brity Copilot deployment reduced the time required to create business operation documents by 75%, while customer service automation rates increased to 60%.[451]

Samsung's vertical integration spans semiconductors (Exynos chips and contract manufacturing), devices (Galaxy smartphones and tablets), cloud services (Samsung Cloud) and enterprise platforms (Samsung SDS).[452] Despite capabilities across the value chain, the company maintains partnerships with external FM providers rather than restricting Galaxy devices to proprietary models or limiting FabriX to Samsung-developed AI.

5.2.3.2 Research innovation: SAIT AI Lab's ultra-efficient models

Beyond Samsung's consumer product deployments and enterprise platform operations, the company's Samsung Advanced Institute of Technology (**SAIT**) AI Lab in Montreal, Canada, demonstrates innovation in FM development challenging conventional assumptions about scale requirements.[453]

As discussed in section 4.5.3, in October 2025, Senior AI Researcher Alexia Jolicoeur-Martineau released the TRM. It was published in a research paper titled "*Less is More: Recursive Reasoning with Tiny Networks*", and claims to have achieved 45% accuracy on the Abstraction and Reasoning Corpus for Artificial General Intelligence (**ARC-AGI**)-1 benchmark (which tests performance on human-like abstract and visual reasoning tasks), compared to Google's Gemini 2.5 Pro (37%), OpenAI's o3-mini-high (34.5%) and DeepSeek-R1 (15.8%).[454] On the more challenging ARC-AGI-2 benchmark, TRM achieved 7.8% accuracy versus Gemini 2.5 Pro's 4.9% and o3-mini-high's 3.0%.[455]

The model's training requirements are much more cost-effective than large language models. According to the author of the paper, training used just 4 Nvidia H100 GPUs and was completed in two days.[456] Jolicoeur-Martineau stated: "*With recursive reasoning, it turns out that 'less is more'. […] A tiny model pretrained from scratch, recursing on itself and updating its answers over time, can achieve a lot without breaking the bank.*"[457]

Samsung released TRM as open-source under an MIT licence on GitHub, enabling anyone to modify and deploy it for commercial applications.[458] This open-source release aligns with Samsung's broader strategy observed in its FabriX platform (distributing competitor models alongside proprietary offerings).

While TRM was designed specifically for structured problems (such as sudoku or mazes) and is not suitable for open-ended text-based or multimodal tasks, the research demonstrates Samsung's continued investment in AI research beyond consumer product deployments.[459] For Samsung's mobile-first consumer devices, ultra-efficient models enabling on-device deployment without cloud connectivity are useful in specific scenarios. These include offline operation in areas with poor connectivity, data privacy, applications requiring immediate responses and providing lower costs.

---

451 The Investor, 'IT solutions provider aims to heighten work efficiency by innovating hyperautomation', 13 September 2023.
452 Samsung Electronics, 'Consolidated Financial Statements for the Year Ended December 31, 2024', 18 February 2025.
Samsung, 'Mobile performance redefined', accessed December 2025.
453 Samsung SAIL, 'Samsung Advanced Institute of Technology AI Lab Montreal', accessed December 2025.
454 Alexia Jolicoeur-Martineau, 'Less is More: Recursive Reasoning with Tiny Networks', 6 October 2025.
455 Ibid.
456 The Outpost, 'Samsung's Tiny AI Model Challenges Industry Giants in Reasoning Tasks', 9 October 2025.
457 Alexia Jolicoeur-Martineau, 'Less is More: Recursive Reasoning with Tiny Networks', 29 September 2025.
458 GitHub, 'Samsung SAIL Montreal / Tiny Recursive Models', accessed November 2025.
459 SiliconANGLE, 'Samsung researchers created a tiny AI model that shames the biggest LLMs in reasoning puzzles', 9 October 2025.

5.2.4 TCS as a service integrator

TCS, India's largest IT services firm, has built an extensive network of partnerships to help their customers adopt AI faster and at scale.[460] TCS has announced or expanded partnerships with several infrastructure providers, including Google Cloud (October 2025), Microsoft Azure (June 2025), AWS (November 2023), IBM watsonx (October 2023), and Nvidia (October 2024).[461] This enables TCS to offer multiple platforms to its customers.

As discussed in the competitive implications section (section 4.4.1), TCS's role as a neutral service integrator expands the GenAI market while increasing competition among FM providers.

Through these partnerships, TCS gains access to competing FMs across all major platforms. AWS Bedrock, Google Cloud Vertex AI and Microsoft Azure offer a selection of models, including those from global developers, such as OpenAI (GPT), Anthropic (Claude) and Meta (Llama), and APAC developers, such as Alibaba (Qwen) and DeepSeek.[462] In addition, each cloud provider offers a larger suite of their proprietary models, with Google Cloud Vertex AI including Google's Gemini model and IBM watsonx including IBM's Granite model.[463]

TCS's operations span 55 countries with over 607,000 consultants, generating USD 29 billion in revenues during the fiscal year ending 31 March 2024.[464] Rather than developing its own AI models, the firm partners with external platform providers to build industry-specific solutions that address enterprise clients' particular operational needs.

TCS trains its workforce across multiple competing platforms. The company is committed to train 25,000 employees on Microsoft Azure OpenAI and to upskill and certify 25,000 employees on AWS generative AI services.[465]

Beyond workforce training on proprietary platforms, TCS built its own platform (TCS AI.WisdomNext™) that runs on top of cloud providers and brings together multiple GenAI services into a single interface.[466] Launched in June 2024, it enables organisations to navigate a quickly evolving AI marketplace and rapidly adopt GenAI technologies at a lower cost by reusing pre-existing components (like built-in guardrails to ensure compliance with local regulations and best practices).[467]

WisdomNext enables real-time experimentation across proprietary, internal, and open-source LLM models, with "intelligent evaluator bots" that compare available models and technology stacks to recommend the best fit for specific workloads.[468] This orchestration approach allows TCS to deploy different FM platforms for different clients based on their needs rather than standardising on any single

---

460 TCS, 'TCS Launches NVIDIA Business Unit to Accelerate AI Adoption for Customers Across Industries', 24 October 2024.

461 TCS, 'TCS Partners with Google Cloud to Integrate Gemini Enterprise for its Workforce and Customers', 14 October 2025.
TCS, 'TCS to Build New AI-Led Solutions for Business Transformation in Collaboration with Microsoft', 20 June 2025.
TCS, TCS launches new Generative AI practice in collaboration with AWS', 27 November 2023.
IBM, 'IBM Launches watsonx Code Assistant, Delivers Generative AI-powered Code Generation Capabilities Built for Enterprise Application Modernization', 26 October 2023.
TCS, 'TCS Launches NVIDIA Business Unit to Accelerate AI Adoption for Customers Across Industries', 24 October 2024.

462 For the list of FMs available in AWS Bedrock, see 'AWS'.
For the list of FMs available in Microsoft Azure, see 'Azure'.
For the list of FMs available in Google Cloud Vertex AI, see 'Google Cloud'.

463 For the list of FMs available in Google Cloud Vertex AI, see 'Google Cloud'.
For the list of FMs available in IBM watsonx, see 'IBM'.

464 TCS, 'TCS Named Global Top Employer for 2025', 18 February 2025.
TCS, 'Record Deal Wins and Robust Margins Mark Strong Finish to TCS' FY24', 12 April 2024.

465 TCS, 'TCS Bets Big on Azure Open AI', 6 July 2023.
TCS, 'TCS and AWS Sign Strategic Agreement to Accelerate Cloud Transformations, Offer Access to GenAI Solutions to Customers', 24 April 2024.

466 TCS, 'TCS AI WisdomNext™', accessed November 2025.

467 TCS, 'TCS Launches WisdomNext™, an industry-first GenAI Aggregation Platform', 7 June 2024.

468 TCS, 'TCS AI WisdomNext™', accessed November 2025.

supplier's offer. The comparison capabilities enable TCS to act as a neutral broker that facilitates model competition rather than channelling enterprises toward any particular supplier.

For client deployments, TCS and AI platforms routinely co-develop industry-specific solutions built on these platforms. In manufacturing, TCS launched a Cognitive Plant Operations Adviser, which integrates with Microsoft Azure to help plant operators achieve factory automation, increase equipment uptime and reduce safety incidents.[469] In retail, the company used Google Cloud capabilities to develop several products which address specific retail challenges:

- TCS Cognitive Visual Receiving uses Vertex AI Vision to automate warehouse receiving processes;
- TCS Talk to Data Insights provides a business intelligence tool for managers, using Vertex AI and CloudSQL to allow for natural language voice queries; and
- TCS Agentic AI-enabled Service Center uses Gemini 1.5 Pro to create an intelligent hub for handling customer queries.[470]

Another retail deployment involved FairPrice Group, one of Singapore's largest retailers, which handles over one million customer interactions daily and has identified 570 distinct types of customer interactions.[471] FairPrice launched its AI-powered Store of Tomorrow in August 2025, with TCS delivering one of its pillars, Grocer Genie, an AI-powered store operations portal using Google Cloud's Vertex AI, BigQuery, and Gemini models.[472] Key features of the system include natural language querying of performance metrics such as sales and inventory, plus automated task creation.

In hospitality, Wyndham Hotels & Resorts, operating approximately 9,100 hotels across 95 countries, extended its partnership with TCS and AWS in November 2023 to manage its digital transformation and use AWS GenAI services including Amazon Bedrock.[473] This partnership enabled Wyndham to migrate its systems to the cloud and standardise its data, building the foundational infrastructure that, according to the company, will support future innovation and enable it to benefit from GenAI services.

5.2.5 Rakuten build and buy strategy

Rakuten Group operates an integrated platform combining over 70 services across e-commerce (Rakuten Ichiba marketplace), fintech (Rakuten Card, Rakuten Bank and Rakuten Pay), telecommunications (Rakuten Mobile), travel, and digital content, serving 1.7 billion members worldwide.[474] This integrated structure, similar to super app platforms prevalent in APAC markets, generates cross-service data that Rakuten uses to develop specialised FMs.

As discussed in the competitive implications (sections 4.2.2 and 4.5.2), Rakuten uses its vertical integration to help develop proprietary FMs while maintaining openness through partnerships with other FM providers.

Rakuten's integrated platform generates proprietary data across multiple service categories. The company operates Rakuten Ichiba (an e-commerce marketplace), Rakuten Card and Rakuten Bank

469 TCS, 'TCS Launches 5G-Enabled Cognitive Plant Operations Adviser to Help Transform Plant Operations', 14 March 2023.
470 TCS, 'TCS Partners with Google Cloud to Accelerate AI-led Transformation in the Retail Sector', 11 April 2025.
471 FairPrice Group, 'FairPrice Group opens Store of Tomorrow at Punggol Digital District', 28 August 2025.
Google Cloud, 'FairPrice Group Unveils 'store of Tomorrow' Program with Google Cloud to Reimagine Its Retail Experiences and Operations', 5 June 2025.
FairPrice Group, 'FairPrice Group', accessed December 2025.
472 Ibid.
473 TCS, 'TCS launches new Generative AI practice in collaboration with AWS', 27 November 2023.
474 Rakuten, 'Rakuten to Collaborate with OpenAI to Develop State-of-the-Art AI Experiences for Consumers and Businesses', 2 August 2023.

(serving 17 million banking customers), Rakuten Mobile (with 9 million subscribers), Rakuten Travel (a travel booking service), and dozens of other services.[475]

Rakuten deployed AI capabilities across its services throughout 2024 and 2025. The company rolled out semantic search, using AI to understand user intent rather than just matching keywords, across 11 services during 2024.[476] On Rakuten Ichiba, the company's e-commerce marketplace, semantic search reduced "zero-hit" search results (that is, searches returning no products) by 98.5%, according to company disclosures.[477] Rakuten attributes a 10.7% year-over-year increase in gross merchandise sales to these search improvements.[478] The company deployed AI-powered product recommendations on Rakuten Ichiba and reported that it improved recommendation efficacy by approximately 60%.[479] Over 30,000 merchants on Rakuten Ichiba use the RMS AI Assistant, a tool helping merchants create product descriptions, marketing content, and customer communications.[480]

The company's AI deployment demonstrates three patterns relevant to understanding FM competition: building proprietary models when it has relevant (Japanese-language) data to do so, relying on multi-homing to choose the best model for each task, and sharing some specialised open-source models available while developing proprietary technology.

Partnership with OpenAI

In August 2023, Rakuten announced a strategic partnership with OpenAI, making Rakuten one of OpenAI's earliest enterprise partners in Asia.[481] Through the partnership, Rakuten announced "Rakuten AI for Business" in November 2023.[482] Rakuten AI for Business supports a diverse set of essential business functions and provides Japanese small and medium enterprises with access to GenAI capabilities powered by OpenAI's technology.[483] The service is optimised for Japanese business culture, regulations, and customs, and requires no technical setup. It includes features addressing Japanese enterprise concerns, such as data security (user inputs not used for training without permission), restricted word management to block confidential information, and job-specific prompt templates based on Rakuten's business use cases.[484] At approximately JPY 1,100 (USD 7) per month, the service targets Japanese SMEs not yet using AI, providing easier and lower-cost access than deploying OpenAI services directly.[485]

Building internal capabilities where strategic advantages exist

While partnering with OpenAI for general-purpose capabilities, Rakuten recognised its proprietary Japanese language data as an advantage for building specialised models. Lee Xiong, Rakuten's Vice Director of AI Research Supervisory Department, explained: "*We have a lot of data, and a lot of the data is in Japanese, in pretty reasonable quality"* and *"the team understands it, so they're able to do better filtering*."[486] The company's multi-stage data filtering and annotation process enables higher-quality

---

475 Rakuten, 'Message To Shareholders and Investors', March 2025.
Rakuten Bank, 'Number of accounts and deposits', accessed November 2025.
Rakuten Mobile, 'Rakuten Mobile Surpasses 9 Million Subscribers', 7 July 2025.
476 Rakuten Today, 'Ting Cai on Rakuten's embrace of AI 2.0', 22 February 2025.
477 Rakuten, 'FY2024 Second Quarter Consolidated Financial Results', 9 August 2024, p 66.
478 Rakuten Today, 'Rakuten AI: Our agentic future starts here', 27 August 2025.
479 Rakuten Today, 'Supercharging Rakuten's merchants with AI: CEO Mikitani', 12 March 2025.
480 Ibid.
481 Rakuten, 'Rakuten to Collaborate with OpenAI to Develop State-of-the-Art AI Experiences for Consumers and Businesses', 2 August 2023.
482 Rakuten, 'Rakuten Selects OpenAI as Strategic AI Partner and Collaborator on New 'Rakuten AI for Business' Platform', 14 November 2023.
483 Rakuten Today, 'Introducing Rakuten AI for Business: A GenAI solution for SMEs in Japan', 4 February 2025.
484 The Fast Mode, 'Rakuten's Japanese-Optimized GenAI Tool Hits the Market,' 29 January 2025.
Rakuten Today, 'Introducing Rakuten AI for Business: A GenAI solution for SMEs in Japan', 4 February 2025.
485 Ibid.
486 Rakuten Today, 'Inside Rakuten AI: Lee Xiong on Japanese LLMs and the future of AI', 23 January 2025.

training for Japanese language tasks compared to generic internet datasets and places Rakuten in a particularly strong position to develop Japanese-specific language models.[487]

In March 2024, Rakuten released Rakuten AI 7B, a 7 billion parameter Japanese language FM achieving top performance among open Japanese LLMs against the Japanese LM Evaluation Harness benchmark at launch.[488] The model was created by continually training Mistral AI's open-source Mistral-7B-v0.1 base model using curated Japanese and English datasets. Rakuten built a custom tokeniser optimised for Japanese characters.[489] Where English-based models represent each Japanese character as a single or even multiple tokens (increasing processing cost), Rakuten's tokeniser can represent multiple Japanese characters as a single token, reducing the token count per request and thereby lowering the computing and usage costs for developers running Japanese-language applications.[490]

In December 2024, Rakuten unveiled Rakuten AI 2.0, an 8x7B MoE model, and Rakuten AI 2.0 mini, a 1.5 billion parameter small language model.[491] MoE uses only the relevant model components for each input rather than processing inputs through the entire model. According to Rakuten, this delivers performance comparable to models eight times larger while consuming approximately four times less computing power during inference.[492] These models were released in February 2025 through Rakuten's Hugging Face repository, an open-source platform for sharing and distributing machine learning models.[493]

With the release of Rakuten AI 7B in March 2024, Rakuten established a multi-homing strategy combining an OpenAI partnership for general-purpose capabilities with proprietary Japanese-optimised models for language-specific tasks. This approach enables performance optimisation: using the best model for each task rather than relying on a single provider for all use cases. The pattern also demonstrates asymmetric sourcing. Rakuten uses OpenAI as a commercial service while building on Mistral's open-source foundation for proprietary development, showing that companies can combine commercial partnerships with reliance on open-source models.

The multi-homing strategy also relies on using different providers for different roles. For example, Rakuten Mobile integrated OpenAI's Realtime API for voice assistants, combining OpenAI's speech-to-speech capabilities with Rakuten AI 7B for Japanese language understanding.[494] Rakuten AI for Business uses OpenAI technology for enterprise customers while consumer-facing services deploy proprietary Japanese-optimised models. Rakuten has not publicly disclosed the distribution of workloads between its proprietary models and OpenAI's capabilities.

Frontier development in agentic AI

Beyond individual service deployments, Rakuten launched Rakuten AI in July 2025 as an agentic AI platform, meaning an AI system that can execute actions on behalf of users rather than just providing information or recommendations.[495] The platform is integrated into the Rakuten Link app (Rakuten Mobile's communication app) and available as a web application, initially supporting Rakuten Ichiba, Rakuten Books, Rakuten Fashion, and Rakuten Rakuma services.[496] According to Rakuten CEO Mickey Mikitani, the platform aims to enable capabilities where "*AI* [is] *booking your trip, arranging your return*

487 Rakuten, 'Rakuten Releases High-Performance Open Large Language Models Optimized for the Japanese Language,' 21 March 2024. Rakuten Today, 'Rakuten's Open LLM Tops Performance Charts in Japanese', 19 April 2024.
488 Ibid.
489 Ibid.
490 Ibid.
491 Rakuten, 'Rakuten Unveils New AI Models Optimized for Japanese', 18 December 2024.
492 Ibid.
493 Rakuten, 'Rakuten AI 2.0 Large Language Model and Small Language Model Optimized for Japanese Now Available', 12 February 2025.
494 Rakuten Today, 'How are AI voice agents changing customer service? Rakuten Mobile collaborates with OpenAI to integrate new technology', 23 October 2025.
495 Rakuten, 'Rakuten Announces Full-Scale Launch of Rakuten AI', 30 July 2025.
496 Ibid.

train, helping you shop – all based on your preferences, profile and previous behavior."[497] Rakuten Chief AI and Data Officer Ting Cai described the vision as empowering users to accomplish tasks "*from complex research to immediate actions,*" with AI agents that "*drive deeper engagement and create even greater value*."[498]

Users can interact with the AI using natural language to perform tasks across Rakuten's ecosystem. For example, a user could ask "find gift recommendations under ¥5,000 for my mother's birthday," and the AI would search Rakuten Ichiba and present options. Alternatively, a user could request "book a hotel in Kyoto for next weekend under ¥15,000," and the AI would search Rakuten Travel and provide options. Rakuten AI's rollout to Rakuten Ichiba was planned for Autumn 2025, initially focusing on personalised product recommendations using integrated specialised agents and insights data that includes user attributes, preferences and purchasing trends.[499] The platform's broader agentic vision includes enabling end-to-end transaction completion (completing purchases and bookings across services), though the company has not publicly disclosed a specific timeline for full transaction execution capabilities.[500]

The agentic capability reflects Rakuten's platform integration advantage. Because Rakuten controls authentication, payment systems, and data across services, the AI can execute transactions, not just suggest them, across shopping, travel, financial services, and telecommunications within a single conversational flow. This differs from standalone chatbots that can only provide information or links to external services. The deployment demonstrates how super app platform structures, prevalent in APAC markets where agentic AI adoption is accelerating, may enable richer agentic capabilities through use of integrated infrastructure.

Despite using its proprietary ecosystem data to build Rakuten AI models, the company released all models under the Apache 2.0 license in February 2025, enabling free commercial use.[501] The models are available on Rakuten's Hugging Face repository for various uses including: text generation, summaries, question answering, dialogue systems designed to have natural conversations with humans, and as foundations for specialised applications.[502]

5.2.6 CBA co-development

CBA is Australia's largest bank, serving over 17 million customers and handling approximately 40% of Australian payment transactions.[503] The bank operates retail, business, and institutional banking across Australia and New Zealand, employing approximately 48,900 staff.[504]

As discussed in the competitive implications (section 4.3), CBA's architecture relies on multi-homing across four major FM providers without the risk of supplier lock-in.

CBA launched its Customer Engagement Engine in 2016, using approximately 1,000 machine learning models and 157 billion data points, establishing AI capabilities well before widespread GenAI adoption.[505] By 2024, CBA operated over 2,000 AI models making approximately 55 million decisions daily, earning rankings as the number one bank in APAC and fifth globally for AI maturity according to the Evident AI

497 Rakuten Today, 'Mickey Mikitani: The age of agentic AI has arrived', 7 August 2025.
498 Rakuten, 'Rakuten Announces Full-Scale Launch of Rakuten AI,' 30 July 2025.
499 Ibid.
500 Rakuten Today, 'Rakuten AI: Our agentic future starts here', 27 August 2025.
501 Rakuten, 'Rakuten AI 2.0 Large Language Model and Small Language Model Optimized for Japanese Now Available', 12 February 2025.
502 Ibid.
503 CBA, '2024 Annual Report', 14 August 2024, p 4.
CBA, 'Commonwealth Bank is at the forefront of real-time sanctions screening', accessed November 2025.
504 CBA, 'Our company', accessed November 2025.
505 CBA, 'CBA announces new reimagined banking services', 24 May 2023.

Index.[506] CBA's established technical sophistication, substantial engineering resources and investment capacity (USD 2.3 billion of technology investment), enables it to manage multiple suppliers.[507]

CBA operates simultaneous co-development partnerships with four major FM providers:

- AWS (infrastructure and AI Factory);
- Microsoft (productivity tools and custom Microsoft Copilot development);
- Anthropic (applications in fraud detection); and
- OpenAI (exploring scam prevention).

Despite partnerships involving joint engineering, roadmaps, and solutions, CBA maintains options to use multiple suppliers through choices that preserve flexibility. CBA has developed internal infrastructure that standardises deployment and enables supplier-specific implementations. In particular:

- The bank's Development Operations Hosting Platform, which has been recently enhanced, provides standardised infrastructure to build and release AI applications.[508]
- CBA's hosting platform is integrated with Amazon Bedrock and Amazon Bedrock Knowledge Bases to access various models from leading AI companies.[509]
- CBA has implemented 11 guardrails on its GenAI models' inputs and outputs to enforce its security and compliance policies.[510]

CBA has collaborated with multiple external providers as discussed below.

CBA x Microsoft

CBA and Microsoft extended their partnership in March 2024 with a focus on GenAI and cybersecurity.[511] A survey of the bank's 10,000 active Microsoft Copilot users reported that 84% would not want to work without it.[512] CBA deployed GitHub Copilot during testing, with engineers accepting nearly 80,000 lines of AI-suggested code, representing approximately one-third of all recommendations made.[513]

Beyond consuming Microsoft's existing products, CBA and Microsoft are co-developing CommBank Copilot, a custom AI assistant designed to help customers resolve banking queries and better understand their financial position.[514] The project involves Microsoft engineers working directly with CBA teams on experiments and use cases.[515] Based on current information, CommBank Copilot remains in development with no deployment having been disclosed.

CBA x AWS

CBA's relationship with AWS spans over a decade. The bank began cloud migration of critical workloads in 2015 when it transferred CommSec's web and mobile from on-premise data centres to AWS and later

---

506 CBA, 'CommBank accelerates AI integration with major data migration to cloud', 4 June 2025.
Evident, 'Here's the 2024 Evident AI Index', 17 October 2024.
507 CBA, '2025 Annual Review', 14 September 2025, p 3.
508 CBA, 'CommBank and AWS expand collaboration to deliver global best cloud and AI capabilities, enabling idea to production in six weeks', 4 February 2025.
509 CBA, 'CommBank and AWS expand collaboration to deliver global best cloud and AI capabilities, enabling idea to production in six weeks', 4 February 2025.
AWS, 'Amazon Bedrock', accessed November 2025.
510 CBA, 'Customer safety, convenience and recognition boosted by early implementation of Gen AI', 28 November 2024.
511 CBA, 'CBA and Microsoft deepens Gen AI partnership', 11 March 2024.
512 Microsoft, 'Commonwealth Bank of Australia invests in AI skills and Microsoft Copilot to drive innovation', 23 June 2025.
513 Microsoft, 'CommBank demonstrates the real-world benefits of AI', 7 February 2024.
514 CBA, 'CBA and Microsoft deepens Gen AI partnership', 11 March 2024.
515 Ibid.

transferred CommSec's applications to AWS in 2019.[516] CBA progressively migrated other core applications platforms including NetBank and MUREX from on-premises environments to AWS.[517]

In September 2024, CBA and AWS jointly started the AI Factory using Amazon EC2 P5 Instances with Nvidia H100 GPUs and Amazon SageMaker.[518] The AI Factory will allow CBA employees to develop, test, and fine-tune LLMs at increased speeds.[519] According to CBA's Chief Data and Analytics Officer Dr Andrew McMullan, the AI Factory accelerates the bank's AI development time by approximately four times compared to previous rates.[520] In February 2025, CBA formalised a five-year strategic collaboration designating AWS as the bank's "preferred cloud provider".[521] In June 2025, CBA completed a full data platform migration to AWS, moving over 61,000 data pipelines, in collaboration with AWS and HCLTech.[522] This "preferred provider" designation demonstrates that formal infrastructure commitments can coexist with a diverse set of application-layer suppliers.

CBA and AWS co-developed the CommBiz GenAI messaging service, deploying it to tens of thousands of business banking customers, taking six weeks from concept to deployment.[523] The solution uses Amazon Bedrock Knowledge Bases to access multiple FMs including Claude 3 (Anthropic) and Cohere's LLMs.[524] The service responds to business customer queries in everyday language by pulling information from over 80 user guides.[525] CBA's use of Bedrock's multi-model capability rather than a single-model API demonstrates it prefers flexibility even when co-developing with AWS.

CBA x Anthropic

CBA announced an expanded strategic partnership and an undisclosed investment in Anthropic in March 2025.[526] The partnership involves CBA engineers working directly with Anthropic's AI experts to develop applications in fraud prevention and customer service.[527] CBA has successfully employed GenAI for fraud detection before and although the bank has not provided details on the GenAI models used, CBA achieved measurable results under this fraud prevention strategy. The bank processes and analyses over 20 million payments daily using GenAI to flag suspicious transactions, sending 20,000 to 35,000 proactive warning alerts per day to customers.[528] The bank reported a 50% reduction in customer scam losses and a 30% drop in customer-reported frauds due to AI-powered features.[529]

CBA x OpenAI

In August 2025, CBA became OpenAI's strategic banking partner in Australia through a multi-year agreement.[530] The partnership provides CBA employees with progressive access to OpenAI's products including ChatGPT Enterprise and involves CBA and OpenAI engineers collaborating to explore GenAI solutions for fraud detection and personalised services.[531] The OpenAI partnership focuses on exploration

---

516 AWS, 'How CommBank made their CommSec trading platform highly available and operationally resilient', 19 August 2025.
517 CBA, 'CommBank and AWS expand collaboration to deliver global best cloud and AI capabilities, enabling idea to production in six weeks', 4 February 2025.
518 CBA, 'CommBank revolutionises banking by activating AI Factory with AWS', 17 September 2024.
519 Ibid.
520 Ibid.
521 CBA, 'CommBank and AWS expand collaboration to deliver global best cloud and AI capabilities, enabling idea to production in six weeks', 4 February 2025.
522 CBA, 'CommBank accelerates AI integration with major data migration to cloud', 4 June 2025.
523 CBA, 'CommBank and AWS expand collaboration to deliver global best cloud and AI capabilities, enabling idea to production in six weeks', 4 February 2025.
524 Ibid.
525 Ibid.
526 CBA, 'CommBank expands strategic partnership with generative AI company, Anthropic', 14 March 2025.
527 Ibid.
528 CBA, 'Customer safety, convenience and recognition boosted by early implementation of Gen AI', 28 November 2024.
529 Ibid.
530 CBA, 'CommBank and OpenAI embark on Australia-first strategic partnership to advance AI solutions', 13 August 2025.
531 Ibid.

and future capability development with no confirmed uses or measurable results. CEO Matt Comyn stated the partnership reflects CBA's "*commitment to bringing world-class capabilities to Australia*".[532]

CBA operates both Anthropic and OpenAI partnerships simultaneously for overlapping fraud and scam detection use cases. The OpenAI partnership represents parallel exploration with another partner and was announced five months after the Anthropic expansion. This parallel development across competing suppliers for the same safety-critical use case demonstrates that CBA actively evaluates different suppliers to enable switching as the need arises.

Other collaborations

CBA uses H2O.ai to power intelligent digital processing of documents, text and image data.[533] It was the first organisation globally to use H2O.ai's Document AI, a tool that enables advanced understanding and interpretation of digitised documents.[534] Over 1,000 employees have received H2O.ai's tools and training, and CBA taught over 170 staff, including non-data scientists, to experiment and build models on the H2O.ai platform.[535] As part of the partnership, CBA is also providing H2O.ai's state-of-the-art AI Cloud to its entire organisation.[536] CBA and H2O.ai also co-developed an AI model capable of identifying digital payment transactions that contain concerning messages in the messaging field in an attempt to reduce technology-facilitated abuse.[537]

CBA also works with Apate.ai to harness near real-time scam intelligence and help protect its customers from scams.[538] Apate.ai deploys thousands of conversational AI bots daily to disrupt scammers via text-based conversations and voice calls, feeding intelligence insights directly into CBA's scam control systems.[539]

Beyond these supplier-specific initiatives, there are several other use cases of GenAI across the bank, with each business function developing an in-house model or adopting external FMs to suit its needs.

CBA's Seattle Technology Hub

In March 2025, CBA established a dedicated Technology Hub in Seattle, five months after opening the AWS AI Factory and one year after expanding the Microsoft partnership.[540] The hub places CBA engineers close to major technology partners (AWS, Microsoft, OpenAI and Anthropic). CBA technologists participated in three-week exchange programmes working alongside teams from AWS, Microsoft, Anthropic, and H2O.ai.[541] Microsoft Corporate VP Charles Lamanna and AWS VP Swami Sivasubramanian both emphasised the hub's role in strengthening existing collaborative work.[542]

CBA Group Chief Information Officer Gavin Munroe acknowledged the multi-homing strategy: "*Working in partnership with Microsoft, alongside other external partners, gives us the opportunity to access the global expertise in a range of areas*."[543]

---

532 Ibid.
533 H2O.ai, 'How Commonwealth Bank is transforming operations with Document AI', 11 April 2023.
534 CBA, 'H2O.ai partnership powers personalised banking and intelligent digital processing', 16 November 2022.
535 Ibid.
536 Ibid.
537 CBA 'In a world first, CBA shares its artificial intelligence model to help reduce technology-facilitated abuse', 8 November 2023.
538 CBA, 'CommBank harnesses near real-time, AI-powered intelligence to outsmart the scammers', 27 June 2025.
539 Ibid.
540 CBA, 'CommBank establishes Seattle Tech Hub to further accelerate its AI capability', 27 March 2025.
541 Ibid.
542 Ibid.
543 CBA, 'CBA and Microsoft deepens Gen AI partnership', 11 March 2024.