MEDIATEK | Counterpoint Technology Market Research

零一万物 01.AI | 阿里云 | 百川智能 BAICHUAN AI | 虎牙直播 HUYA.COM | KUGOU | oppo | Soul APP | Tencent AI Lab | 腾讯混元 | vivo

Generative AI Phone Industry Whitepaper

May 2024

This whitepaper has been jointly released by Counterpoint Research and MediaTek along with other partners, including 01.AI, Alibaba Cloud Tongyi Qianwen, Baichuan-AI, Huya, Kugou, OPPO, Soul, Tencent AI Lab, Tencent Hunyuan and vivo (in alphabetical order).

Table of Contents

Preface

Following the launch of ChatGPT on November 30, 2022, large language models (LLMs) quickly gained popularity. The generative AI revolution initially sparked by ChatGPT amazed global consumers with the novel human-machine interaction experience, while also making the industry fully realize the enormous potential of generative AI in consumer (or To Consumer business, To C) segments. Smartphones, as the most important personal intelligent devices, with a global user base of over 4 billion, undoubtedly play a crucial role in determining the success of generative AI-to-consumer applications.

On the other hand, the smartphone industry, upon entering the 5G era, also needs a truly disruptive innovation to give consumers smarter, personalized and safer experiences. This will transform the phone into a personal intelligent agent and a powerful mobile productivity vehicle, laying a solid foundation for smartphone development over the next decade. The integration of generative AI technologies and smartphones perfectly meets such demand, as it will empower the smartphone industry in all aspects, revolutionizing both the hardware and software, as well as the related mobile internet content ecosystem.

The combination of generative AI and smartphones is both a necessity for the whole industry and a crucial path to the democratization of AI. Leveraging smartphones as a medium, global phone users will be able to enjoy the benefits of the fast-developing generative AI technology, more conveniently and more efficiently.

This whitepaper proposes the concept of generative AI phones and discusses the strategies of stakeholders across the generative AI phone industry including chip vendors, smartphone OEMs, large model providers and APP developers. The whitepaper also discusses the software and hardware landscape of generative AI phones and makes related forecasts through 2027.

Chapter 1: The Onset of New Generative AI Phone Era

A new generation of flagship smartphones arrived in the market between the end of 2023 and Q1 2024 – smartphones supporting an increasing number of generative AI capabilities. Smartphone OEMs and their ecosystem partners are actively embracing the generative AI trend, while also actively exploring various possibilities, focusing on creating high-frequency use scenarios that are valuable to users. This process is expected to continue throughout 2024, which Counterpoint believes will be the inaugural year of generative AI phones.

The History of Handset Smartization

About two decades ago, operating systems represented by Nokia Symbian first allowed users to download apps and use apps as gateways to access services and digital content. The model has persisted to this day. This shift enabled developers worldwide to join the smartphone industry, offering a rich variety of app choices to users and fostering the vigorous development of the mobile internet ecosystem. Smartphones have gradually become indispensable as they can connect to services for leisure and entertainment, communication and social, health and travel services, shopping and mobile office work.

The launch of the iPhone in 2007 revolutionized traditional phone design philosophy. The physical keyboard phased out and the touch screen became the core of human-machine interaction. However, as time passed, touch input became increasingly inefficient in some scenarios, often requiring multiple user interventions to get the service. In this context, new interaction modes, including intelligent voice assistants, gestures and eye tracking, have been invented to create a smoother, more efficient, and more user-friendly experience.

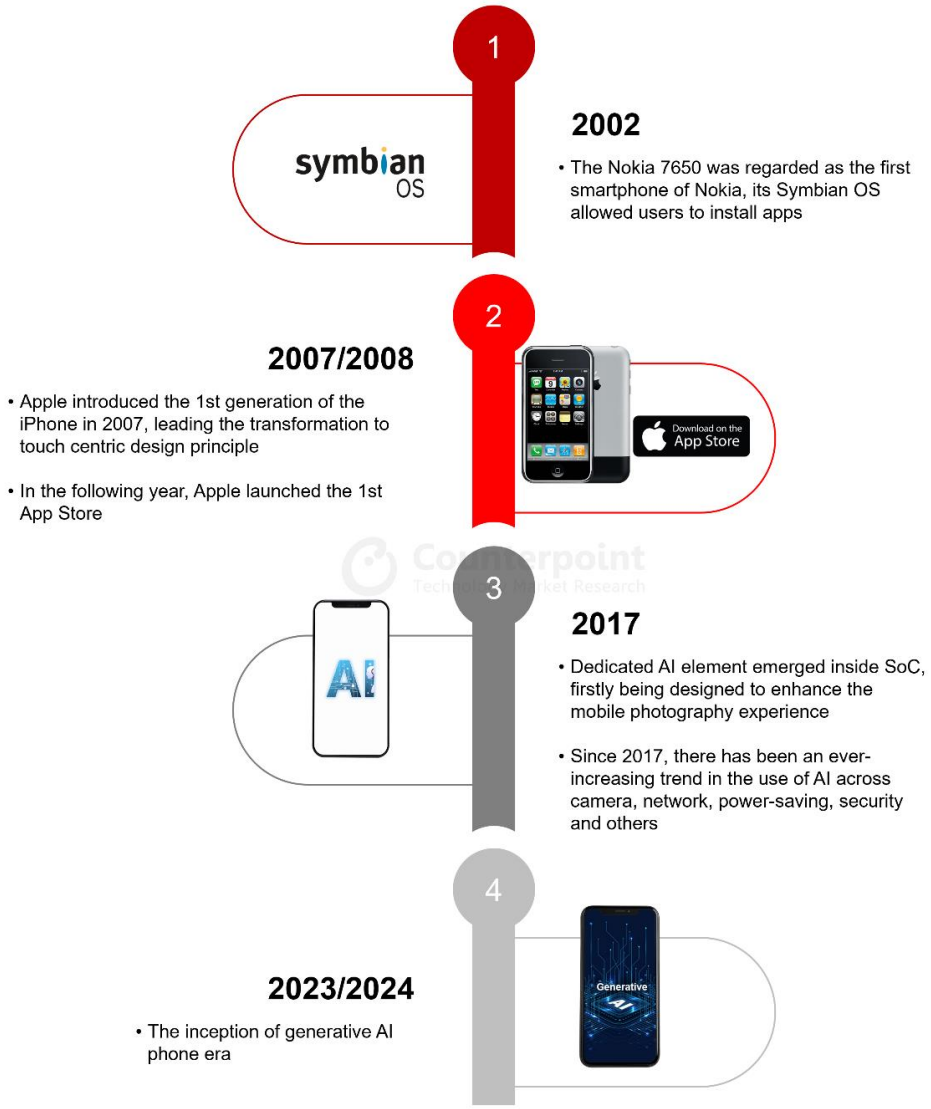
Revolutionizing smartphones with AI technology could date back to 2017 when Apple released the first rear-dual-camera iPhone 7 Plus. Also, the Android camp began to incorporate independent NPU computing units into their SoC platforms for running deep learning models related to photography enhancement. Since then, AI technology has gradually been used by smartphone OEMs in more areas, such as enhancing security, optimizing battery life, and improving network performance. However, computational photography remains the primary application field of AI. The introduction of LLMs into smartphones marked the transition from the implementation of small and medium-sized models to the deployment of large models for smartphones.

Thanks to the capability of generative AI large models, smartphones are set to undergo a new round of innovation. Firstly, in terms of human-computer interaction, with the support of LLMs, new multimodal interactions will replace the traditional interaction mode centering around touchscreen. The UI mode would gradually transit from Graphic User Interfaces (GUI) to Voice User Interfaces (VUI). Users will be able to communicate with their smartphones in a more intuitive and natural way. Secondly, multimodal input and output capabilities can significantly enhance smartphones' productivity. They can generate charts, texts,

music, images and even videos based on various forms of input information from users. The inputted images and videos can also be edited.

Finally, as the integration deepens, generative AI technology will give rise to at least one AI agent on smartphones. Each smartphone could have one or more AI agents. Centered around the user, these AI agents will continuously learn from the user's behavior and habits, intelligently recognize the user's intentions, and timely recommend personalized content and services to the user. Counterpoint expects that these agents would become exclusive gateways for each user to get services. However, it is expected that they will coexist with apps for a long time.

Exhibit 1 The History of Handset Smartization



Source: Counterpoint Research

Counterpoint's Definition of Generative AI Phones

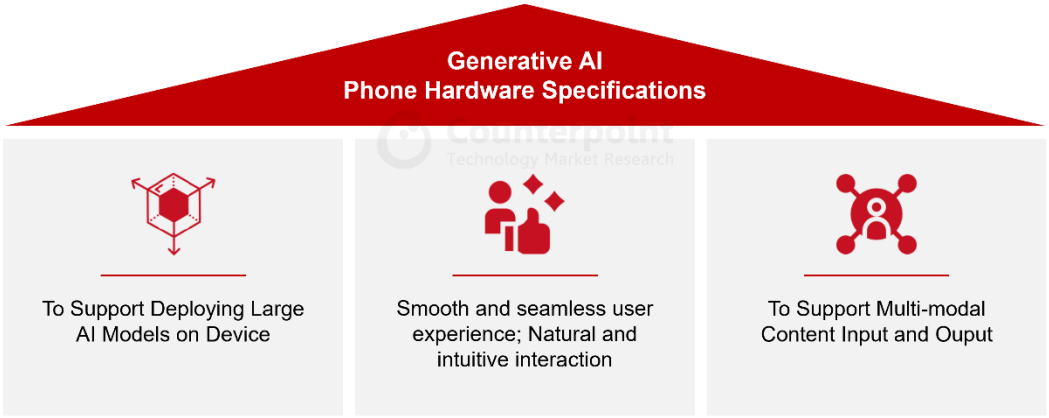
Counterpoint believes that the integration of generative AI with smartphones will undoubtedly trigger a profound transformation. Reflecting on every technological innovation in the past, in the early stages of exploration, new functions and features will first be introduced to flagship and sub-flagship products which have more abundant computing resources, and then quickly become an important differentiated selling point. Over time, the capabilities of generative AI will accelerate to move down price bands and benefit a broader consumer base worldwide.

Hence, Counterpoint has proposed the concept of generative AI phones and its definition in light of the current status of generative AI applications and the future development and evolution of generative AI phones.

Generative AI phones use large-scale, pre-trained generative AI models to achieve multimodal content generation, contextual awareness, and demonstrate increasing human-like capabilities over time. The advent of generative AI marks the onset of a new era in smartphone development. In the long term, smartphones will develop into mobile intelligence. To be considered a generative AI phone, a device must meet the following criteria:

- On-device or hybrid processing of foundational models, which enables the smartphone to perform complex generative AI tasks without relying solely on cloud-based servers.
- Multimodal capabilities, allowing the device to process text, images, voice and other forms of input to generate a variety of outputs, with use cases such as translations, image generation and video generation.
- A user experience that is fluid and seamless, enabling the device to provide natural interactions and intuitive responses to the user's inputs and requests.
- Hardware specifications that are capable of supporting the above capabilities, include but are not limited to a processor built on the latest process node designed to undertake AI workload leveraging integrated or standalone neural network processing units (such as APU/NPU/TPU), large memory capacity with high bandwidth, stable and high-speed connectivity options, and hardware-level and software-level security solutions.

Exhibit 2 Counterpoint's Definition of Generative AI Phone



Source: Counterpoint Research

2024 marks the breakout year for generative AI phones. With cooperation across the industry chain, leading Android OEMs have successfully deployed LLMs with 7 billion parameters on smartphones. Some capabilities have also been provided to APP developers. Representative capabilities include image and video generation based on Diffusion models (normally at low resolution for on-device AI, 480p for instance), natural language processing based on LLMs, such as voice-to-text, text-to-voice, human-machine dialogue for specific tasks, real-time translation, and information Q&A, along with text proofreading, text generation, paraphrasing, and text summary based on sLLM models (lightweight language models).

Device+Cloud Hybrid To be Mainstream in Deploying Generative AI

Compared to the limited computing and memory resources on mobile phones, cloud servers undoubtedly possess more computing power to operate and train larger-scale AI models. Currently, some complex generative AI tasks are mainly performed by large models on the cloud side. Looking ahead, Counterpoint believes that the hybrid implementation of on-device and cloud large AI models will be the mainstream solution for generative AI phones. On one hand, in the coming years, on-device large AI models will continue to grow in both size and efficiency, meaning that users can access most services powered by local generative AI capabilities. Large on-device AI models will also provide data desensitization, compression and other data preprocessing for tasks that require the cloud's support. This would help protect user privacy. On the other hand, cloud-side large AI models can bring more value-added services to users, such as providing higher-quality content output including film and animation production. Cloud-side large AI models can also help build intelligent collaboration platforms namely cloud offices that break the boundaries of physical space, allowing massive numbers of people to participate in the same project at the same time.

In summary, generative AI phones should be viewed from the perspective of development. The generative AI use cases currently showcased may only be the tip of the iceberg. However, no matter at which stage

of development, device-cloud collaboration can meet the needs of all parties in the industry while maximizing the use of computing resources distributed on both sides. On the device side, every upgrade of the mobile computing platform could give rise to a significant breakthrough in mobile AI computing power. Accordingly, consumers can expect a smoother and richer generative AI experience. At the same time, the multimodal capabilities of on-device generative AI will also be further strengthened. Counterpoint expects that multimodal capabilities, including multimodal input and output, are one of the keys to realizing the vision of generative AI phones. They are also the basis for realizing diversified modes of user interaction.

Multi-modal Capacity Key to Realize Vision of AI Agents

As mentioned earlier, the multi-modal capability is key to realizing the vision of generative AI phones. In Counterpoint's forecast period, as generative AI phones develop, the following innovations can be expected:

First, interaction modes with smartphones will become more intuitive and diversified. On one hand, large AI models are required to recognize and understand different types of input content from users, which could be text, a voice clip, a spreadsheet, an image or a video. On the other hand, large AI models with multi-modal output capability could be more user centric. It can choose the most suitable way or format to generate content based on the context and use scenario. Multi-modal interaction is the key to unlocking a new interactive experience.

For a long time, the voice assistant function was considered not very helpful in everyday applications. But with multimodal LLMs, the voice assistant function will become more intelligent. It will be able to more accurately recognize and understand human natural language, not only quickly understanding and responding to user commands but also possessing abilities like voice-to-text conversion and multi-turn conversations. In an increasing number of scenarios, more natural voice interaction will become the preferred method for interacting with smartphones.

Secondly, the advent of Sora earlier this year and its ability to generate videos have attracted significant industry attention. Sora has become the new benchmark of multimodal AI applications. For smartphone users, short video streaming is undoubtedly one of the most popular mobile internet applications at present. Looking forward, video will remain the primary media for sharing digital content. The demand for high-quality, personalized and diversified short video content will continue to grow. Short video production will be one of the most important applications for generative AI. Generative AI phones in the future will understand video input and assist in intelligent video editing, video style change, automatic copywriting, and dubbing generation. In the end, it could become a valuable assistant for video creators. Benefiting from the increase in smartphone AI computing power, language-visual models (LVMs) will also be deployed on smartphones soon after, which it will be possible to generate and edit short videos through

text and voice prompts, greatly enhancing video production efficiency and further promoting the prosperity of the short video ecosystem.

Finally, AI agents will come into reality. Multi-modal large AI models can perceive various types of data simultaneously, including images, text and speech, thereby accessing massive and multi-dimensional data for AI agent training. This means AI agents, like humans, can acquire knowledge from different media, continuously enhancing their understanding of the complex real world. In this process, AI agents will learn human-like skills such as "listening, speaking, reading and writing."

Moreover, with the development of AI cognition, AI agents will acquire more comprehensive capabilities. MediaTek believes that future agents will also possess autonomous decision-making abilities, capable of "Planning", "Memory", and "Action":

- Planning – Having the ability to decompose tasks and self-reflect.
- Memory – Having short-term and long-term memory.
- Action – Directly executing or using tools to complete specific tasks for the user.

According to MediaTek's vision, AI agents could learn and remember smartphone users' habits, interests, and preferences, and then assist them in managing daily life (food, clothing, housing, and transportation) and work (planning, execution, reporting) tasks, allowing humans to focus on setting targets and making decisions. Taking the travel scenario as an example, an AI agent could tailor travel plans for users and dynamically adjust plans timely based on weather, traffic, tourist site information, as well as unexpected situations, making travel easier and more personalized.

The Advantages of On-device Large AI Models

As mentioned earlier, Counterpoint Research believes the device-cloud hybrid mode will be the main method of deploying generative AI technology on smartphones. However, at the same time, Counterpoint Research recognizes the following advantages of deploying large AI models locally:

- Low Latency: In many generative AI task scenarios, network transmission delay is the main reason users cannot enjoy a smooth experience. On-device large AI models can respond to user demands faster, ensuring latency of mere seconds or even milliseconds.
- Security and Privacy: On-device large AI models ensure that personal user data does not leave the smartphone. Combined with the hardware security protection solutions provided by chip companies, this can protect user data and privacy to the highest extent.

- Reduced Dependency on Networks: Large AI models can greatly reduce dependency on networks. Even in areas with weak network or even without access to Internet, the smartphone can still provide necessary generative AI capabilities, offering uninterrupted services to users.

- Personalization: Large AI model on device with the ability to self-learn can grow to be an AI agent tailored for each user, with the AI agent having the ability to provide personalized services and recommendations.

- Reduced Workload on Infrastructure: There are over 4 billion smartphone users globally. As generative AI applications become widespread on smartphones, the demand for AI computing power will at the same time grow exponentially. Deploying large AI models on device makes it possible to complete as many generative AI tasks on device as possible. It could reduce the occupation of network bandwidth and can also significantly reduce workload that consume cloud computing resources.

Chapter Two: Generative AI Phone Ecosystem

Current Status of LLM and Forecast

Generative AI innovation has been sweeping the globe. As major innovators and contributors, Google and Meta's LLM strategies cover various use scenarios from cloud to edge, and then to the device side as well.

In 2017, Google first introduced the Transformer architecture and applied it to natural language processing, making it one of the first leading internet companies to dive into the development of LLM models. In April 2022, Google launched the LLM PaLM (Pathways Language Model), which had 540 billion parameters. In the following year, the upgraded PaLM2 was launched. By adopting Compute Optimal Large Model technology, Google optimized the model's size and enhanced its multilingual, reasoning, and programming capabilities, thus improving PaLM2's performance in real-world applications. By the end of 2023, with the release of the new multimodal model Gemini, Google accelerated the commercialization of LLMs. Counterpoint predicts that Google will prioritize using Gemini to strengthen its services, such as search, Chrome browser and YouTube, and seek to integrate Gemini with the broad Android ecosystem. Currently, Gemini has three versions – Gemini Nano (including Nano-1 with 1.8B and Nano-2 with 3.25B parameters) for device-side applications, and Gemini Pro and Gemini Ultra for cloud-side applications.

In February 2023, Meta released its in-house foundational LLM, called LLaMA. It has various versions including 7 billion, 13 billion, 33 billion and 65 billion parameters. In July 2023, Meta announced to open source of the latest version, LLaMA2. It also reached an agreement with Microsoft, which will support LLaMA2 on Azure and Windows platforms. These moves have helped LLaMA2 gain wide support from industry partners, allowing developers to customize and develop large models atop LLaMA2 for individual business scenarios, significantly advancing the commercialization of LLaMA2. In smartphones, some OEMs have already considered Meta and Google as important partners for deploying generative AI internationally. Meta launched the latest LLaMA3 in 2024 after achieving success with the previous LLaMA2. LLaMA3 has made multiple optimizations and improvements over the previous version, including increased model parameters (from 7B to 8B) and context length (from 4K to 8K), enhanced algorithmic efficiency, and a more refined model training process. Consequently, LLaMA3 has further enhanced the accuracy and generation capabilities of natural language processing. With LLaMA3, Meta will consolidate its leadership in the field of generative AI models,and provides developers and industry partners with more powerful and flexible tools to create more possibilities in their respective business scenarios.

In China, many domestic LLM models have emerged in the past two years, many of which are outstanding. These models' comprehensive performance continues to benchmark the advanced level of international players, and their performance in the Chinese language could be labeled as the most cutting edge in the industry.

01.AI Yi Model Series

01.AI is a global company committed to developing cutting-edge large AI model technology and software applications for the AI 2.0 era. The core of its platform business focuses on constructing industry-leading large AI models, including the open-source Yi-34B, Yi-6B, and the multimodal model Yi-VL, as well as fine-tuned models Yi-34B-Chat (0205), long-text Yi-34B-Chat-200K, and Yi-VL-Plus. The Yi series models support multilingual conversations including English and Chinese and image recognition capabilities for chat. The model Yi-Plus with hundreds of billion parameters is set to debut soon. 01.AI has deep expertise in model miniaturization technologies. Technologies for model miniaturization including pruning, distillation, and quantization have always been a focus of the company's research and development. Additionally, 01.AI plans to gradually release a complete set of open platform middleware and developer tools, to help develop consumer and business applications based on the Yi series' "edge-side models + cloud-side big models" combination. For the consumer-level application business, 01.AI focuses on developing new types of personal efficiency software "Wanzhi" and social functions. For business-level applications, the company actively collaborates with corporate clients to explore business-level "To Business" application scenarios.

Tongyi Qianwen by Alibaba Cloud

Tongyi Qianwen is an LLM developed by Alibaba Cloud Tongyi Lab. The LLM was launched in April 2023 and invited users to test the experience. In September, it was among the first batch for government registration, following which it was officially open to the public. In October 2023, the Tongyi Qianwen 2.0 was launched. Its model parameters reached more than 100 billion. The model's capabilities in complex instruction understanding, literary creation, general mathematics, knowledge memory, hallucination resistance, and more, have been significantly improved, coming to the forefront of the industry. Tongyi large model family has launched PC and mobile applications that can provide users with various supports in life, work, study, entertainment and so on. Tongyi is also one of the most influential open-source large AI models in the industry. The Tongyi large AI model family has made a total of 7 models open-sourced with sizes of 0.5B, 1.8B, 4B, 7B, 14B、32B and 72B parameters, including Base version and Chat version, as well as visual understanding large AI model Qwen-VL and audio understanding model Qwen-Audio. So, it is the first one in the industry to make various sizes of multimodal models open-sourced. The Tongyi Qianwen open-source series models have been downloaded more than 3 million times.

Baichuan AI

The Baichuan first launched a large AI model in June 2023. Its first AI product is Baichuan-7B, which has 7 billion parameters, mainly composed of Chinese pinyin, Chinese characters and words, and English words. It is used for natural language processing, machine translation and Q&A in Chinese and English. Just three months later, Baichuan AI announced to open source its second-generation basic model Baichuan 2, including the 7 billion Baichuan2-7B and the 13 billion Baichuan2-13B. The training corpus of both reached 2.6TB and supports up to 4K context inputs. The models have excellent performance in scenarios such as document generation and multi-round dialogue. With the release of Baichuan 3, the

parameter scale of the Baichuan large model has exceeded 100 billion, the basic general capabilities have been comprehensively improved, and the logical reasoning and semantic understanding capabilities have been significantly enhanced.

Tecent Hunyuan

Tencent's "Hunyuan" large AI model architecture has been upgraded to the Mixture of Expert (MoE) structure, with about a trillion parameters, and excelling at handling complex and multi-tasking use scenarios. In terms of performance in Chinese, it is at the industry-leading level, especially in performing tasks of mathematics, coding, logical reasoning, and multi-turn dialogue. Additionally, Tencent Hunyuan offers models of various sizes to accommodate applications requiring low cost and high inference performance. As a large AI model entirely developed in-house by Tencent, Hunyuan has accumulated industry-leading capabilities through continuous iteration and practice, earning recognition from multiple parties. In the 2023 Science and Technology Award by the China Institute of Electronics, Tencent's "Key Technologies and Applications of the Angel Machine Learning Platform for Large-Scale Data" won the first prize for technological progress.

Building on a solid foundation, Tencent Hunyuan's large AI model is actively pushing the deployment of related applications to create more value. Currently, over 400 business and use scenarios at Tencent have adopted the Hunyuan AI model for testing, and generative AI functions have been available on Tencent WeCom, Tencent Meeting, and Tencent Docs. Tencent Advertising has launched an AI creative advertising platform called Miaosi, based on the Hunyuan model, significantly improving the production and delivery efficiency for advertisers. Furthermore, in collaboration with ecosystem partners, Tencent is offering more than 50 industry solutions powered by large AI model technology for over 20 industries. Tencent is also working with mobile industry partners for various possibilities of using generative AI technology to bring new services and experiences to smartphone users.

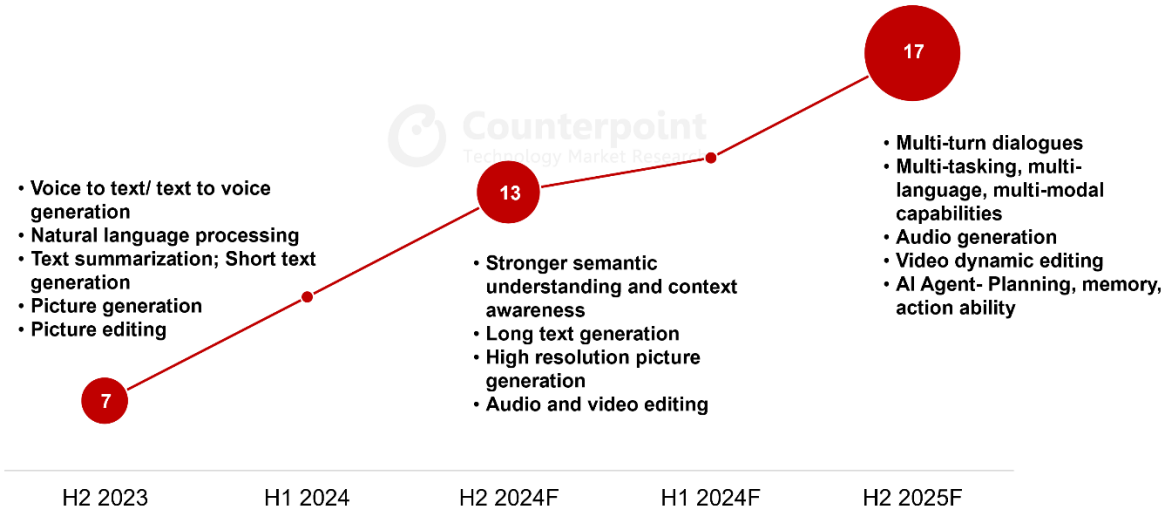
Benefiting from numerous open-source foundational large AI models, developers can customize application models tailored to their business needs, thereby accelerating the deployment of LLMs on both the device side and the cloud side. Moreover, open-source models allow chip vendors and smartphone OEMs work on deep adaption and optimization to run the models. For instance, MediaTek has completed the underlying optimization for models such as Google Gemini Nano, Meta LLaMA2, LLaMA3, Baichuan large AI model, and Tongyi Qianwen on the Dimensity 9300 and 8300 platforms.

Large AI Model Size Projected to Grow in Next Two Years

Currently, several Android flagship smartphones including the vivo X100 series, OPPO Find X7 series, and HONOR Magic 6 series have successfully implemented 7 billion LLMs on the device side. The AI computing power is expected to be a top priority for flagship SoC upgrades in the next two years, enabling local deployment of larger-sized models. Counterpoint believes that the ceiling of parameters of on-device AI models will grow to 13 billion in 2024 and to 17 billion by the end of 2025.

Exhibit 3 On-device Large Model Parameter Forecast, H2 2023-2025(F)

In Billion



Source: Counterpoint Research

The parameter growth of large AI models will further expand the capabilities of generative AI in smartphones, such as more powerful audio and video processing capabilities, as well as support for multiple languages. Some functions for work and creators originally developed for AI PCs will also be available on smartphones. On the other hand, deploying larger AI models generally means higher hardware costs for smartphones and shorter battery life. Considering the limitations of computing power, memory, and heating limits, the smartphone needs to adopt a development strategy different from the cloud, and not blindly pursue a larger parameter scale.

In light of this, chip vendors, such as MediaTek, are exploring new technologies to overcome the limitations of a single large AI model. This includes adopting the Mixture of Experts (MoE) architecture, which integrates multiple expert models. This architecture can dynamically select different expert AI models for different tasks and can also fine-tune each expert model, thereby significantly improving overall performance. With MoE, it may be possible to break the limits relating to a single AI model and match the accuracy of AI models with tens or hundreds of billions of parameters.

Cloud-based large AI models will continue to follow the Scaling Law, namely the more parameters an AI model has, the higher generality and accuracy of the AI model features. In the next two years, more and more cloud-based large AI models will reach the scale of a trillion parameters.

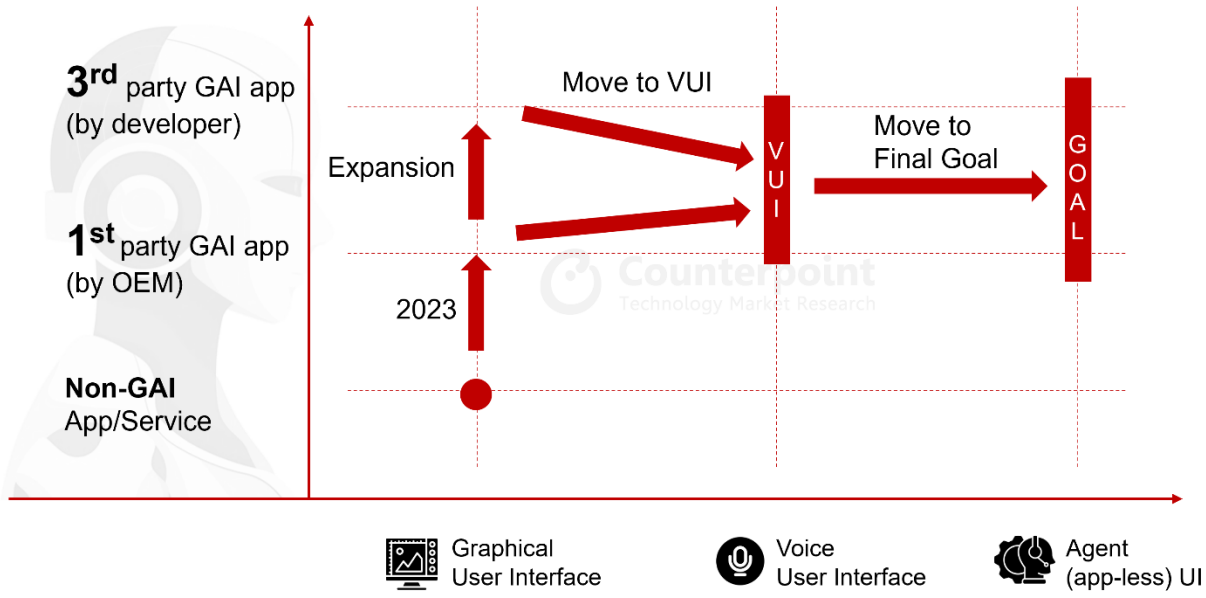
App-based UI and AI Agent UI Coexist for Years

In the era of generative AI phones, collaboration among all parties within the ecosystem is still necessary to drive innovation in products and services by leveraging their respective areas of expertise. Mobile internet companies and app developers are expected to continue leading the development of the mobile content ecosystem, and apps will remain a crucial component of the smartphone user interface.

However, as more smartphone OEMs adopt the user-centric operating system design philosophy, smartphones will be able to automatically generate personalized interface designs based on use scenarios, including automatically adjusting icon layouts, color themes and font sizes.

On the other hand, the diverse and natural interaction experiences offered by AI agents will be favored by smartphone users. As generative AI phones evolve and the generative AI application ecosystem flourishes, more and more resulting functionalities and services will be linked with the AI agent. On this basis, AI agents will renovate the smartphone interaction experience, which means that the traditional Graphical User Interface (GUI) will develop into a Voice User Interface (VUI) and then finally become the brand-new Agent (app-less) UI. This means there would be more interaction between the AI agent and the user and the presence of the app would be softened.

Exhibit 4 Use Interface Development in the Future



Sources: MediaTek, Counterpoint Research

AI agents will gradually become gateways to the digital ecosystem. Users would only need to input their desired services (via voice, text, and more), and the AI agent will directly navigate the user to the desired service page. Additionally, based on an understanding of user habits and current scenarios, AI agents will provide services in a more secure and personalized manner, even directly performing tasks. This is what traditional smartphone applications-centered interfaces cannot realize. But the above transition cannot be realized overnight.

For a long time, both AI agents and apps will coexist in generative AI phones. Moreover, leading internet companies may develop specialized AI agents based on their business ecosystems so that multiple AI agents may coexist and interact in one generative AI phone. They work together to offer intelligent experiences to users.

Generative AI Strategies of Chip Vendors

Chip vendors had already started incorporating independent neural network processing units (NPU) into their SoCs for commercial use by as early as 2017. As of now the AI computing power in a smartphone has increased nearly 20x since then. But running multimodal large AI models locally demands even higher computing power. In the coming years, the AI computing power inside SoCs is expected to maintain a double-digit percentage annual growth rate as semiconductor foundries could migrate to more advanced processes (2nm, 1nm). Chip and IP vendors would also develop new architectures that are more suitable for processing generative large AI models on devices.

Leading chip vendors, such as MediaTek and Qualcomm, have already moved ahead in embracing generative AI phones revolution. They have launched several mobile computing platforms that support the deployment of multimodal large AI models on smartphones. These chipsets include MediaTek’s Dimensity 9300 and Dimensity 8300, Qualcomm’s Snapdragon 8 Gen3 and Samsung’s Exynos 2400, released in Q4 2023, as well as Snapdragon 8s Gen3 and Dimensity 9300+ in Q2 2024, paving the way for the large-scale commercialization of generative AI phones in 2024.

Furthermore, MediaTek and Qualcomm, with their rich experience over the years in AI applications for smartphones, have provided OEMs and developers with a complete set of toolkits to facilitate application development. These are represented by the MediaTek NeuroPilot SDK along with the Qualcomm AI Stack and Qualcomm AI Hub. There are a large number of pre-trained AI models and supporting tools to help developers quickly integrate these models into their applications.

There are three key aspects of MediaTek's generative AI strategy – hardware, software and ecosystem. For the hardware, the new APU engine is built on the mainstream Transformer architecture and supports various integer and floating-point data types, including INT4, INT8, INT16 and FP16, providing industry-leading energy efficiency. In terms of software, MediaTek provides a complete development toolkit and integrated development environment for developers to deploy large AI models and develop AI applications on devices powered by the Dimensity platform. In terms of ecosystem collaboration, MediaTek

continuously expands cooperation with smartphone OEMs, large AI model developers and application developers to jointly build a generative AI mobile application ecosystem.

In summary, leading chip vendors will play a more important role in the era of generative AI phones and will help build a full-stack solution from chipset design to AI models and the toolkit.

Generative AI Strategies of Smartphone OEMs

OEMs are also crucial drivers in the generative AI phones revolution. On one hand, OEMs aim to comprehensively upgrade existing products through generative AI technology, creating differentiated products and continuing to innovate smartphone functions. On the other hand, by leveraging the opportunity brought by the evolution of future AI agents, OEMs can further elevate brands' technological connotation and increase user stickiness to establish long-term and close ties with users.

Many leading OEMs have already released their self-developed large AI models. vivo launched the BlueLM, with the smallest size having 1 billion parameters and the largest one having 175 billion parameters for cloud deployment. vivo as of now has successfully deployed AI models with 1 billion parameters and 7 billion parameters on device and has conducted a run-through of an AI model with 13 billion parameters on device. OPPO has the AndesGPT and its tiny version comes with 7 billion parameters, which was designed for on-device deployment, while the Titan version includes 180 billion parameters. Xiaomi's MiLM is also designed for deployment on device and currently has up to 13 billion parameters. Meanwhile, Samsung and HONOR have each announced multimodal large AI models with 10 billion and 7 billion parameters, respectively.

OPPO's Generative AI Strategy

OPPO has been investing in AI technology for quite some time now. By the end of 2023, it filed over 3,160 AI-related patents globally, primarily in areas of computer vision, speech technology, natural language processing and machine learning. Building on this foundation, OPPO has successfully implemented over 100 AI features, including the AI removal function based on a visual model (average daily usage reached 15 times), and the brand-new XiaoBu Assistant supported by LLM. To better embrace the generative AI phone era, OPPO has established the Binhai Bay Data Center in Shenzhen, with training capabilities powered by thousands of AI computing boards. OPPO established an AI center in January 2024, signalling its strong commitment to AI. In March 2024, the OPPO Find X7 series successfully passed the AI smartphone test by China Telecommunications Technology Labs, making OPPO one of the first OEMs to receive this certification.

In terms of developing its own LLMs, OPPO released a pre-trained language model based on the BERT architecture in 2020. In 2023, the company launched the AndesGPT which is self-trained by OPPO. In the SuperCLUE's top ten basic capabilities list, AndesGPT came to the first place among Chinese large AI models in the "Knowledge and Encyclopedia" capability. OPPO at the same time has developed AI models for both device and cloud deployments, creating a framework of device-cloud collaboration. OPPO's on-

device AI models currently focus on sizes of less than 13 billion parameters, while its AI models for cloud deployment include versions with 13 billion, 34 billion, and over 70 billion parameters. This framework enables flexible adaptation to different business scenarios.

On the device side, OPPO continuously adds new capabilities to its XiaoBu assistant, including speech drafting, brainstorming, document assistance and AI painting functions, to enhance productivity. Additionally, OPPO focuses on creating AI experiences for high-frequency scenarios such as image generation, phone call content summarization and real-time translation. On the cloud side, OPPO aims to realize comprehensive knowledge search and knowledge Q&A through partnerships with third-party platforms and help users shorten the path to acquiring knowledge and services. OPPO has launched several related functions including XiaoBu English Teacher, XiaoBu Interviewer and XiaoBu Q&A.

Driven by the philosophy of 'for user value' and 'win-win collaboration,' OPPO works closely with chip vendors, leading domestic and international large AI model technology providers, and ecosystem partners by adopting MoE architecture. Its self-developed 7-billion-parameter LLM achieves industry-leading inference speed on the MediaTek Dimensity 9300 platform, and its 1-billion-parameter visual model can generate images (at 512x512 resolution) within seconds.

vivo's Generative AI Strategy

vivo is one of the earliest smartphone OEMs to establish an AI team. It formed its AI team in 2017 and set up a Global AI Research Institute in 2018, maintaining a team of more than a thousand experts. In November 2023, the company released its self-developed AI large model, BlueLM, and was among the first batch in the industry to deploy large AI models on smartphone and realize device-cloud hybrid collaboration.

The advantages of vivo BlueLM:

- Massive Data: Since establishing its knowledge graph team in 2018, vivo has accumulated 18,000T of multimodal data and 6,750T of high-quality Chinese text data, including over 30T of training data. This is equivalent to five times of China National Library's collections.
- Comprehensive Model Matrix: For different use scenarios, computing capabilities, and cost requirements, vivo offers the AI model matrix with different sizes ranging from billions, tens of billions to hundreds of billions of parameters. This ensures better experiences by offering AI models from small to big size ones. In 2024, vivo will significantly upgrade its large vision model and plans to release large audio and multimodal AI models.
- Powerful Computing Resource: Currently, vivo possesses abundant training capabilities. It supports its self-developed large AI models with powerful computing and engineering capabilities.
- Efficient Algorithms: vivo has published over 70 papers at top AI academic conferences (including AAAI, ICLR, ECCV, CVPR and InterSpeech) and has applied for more than 800 invention patents.

- Support various use Scenarios: vivo’s AI models cover various application scenarios, including AI-enhanced imaging, office work, video, music, gaming, accessibility, lifestyle, health, travel and security.
- With continuous technology accumulation in algorithms, engineering, data, systems and applications, vivo's BlueLM large AI models have achieved industry-leading performance in the Chinese language. In the Trusted AI Large Model Standard Compliance Evaluation organized by the China Academy of Information and Communications Technology, the BlueLM received the highest rating of 4 stars+ in the large language model capability assessment.

Powered by the BlueLM, the AI assistant Blue Heart V / Blue Heart Qianxun is capable of providing efficient, intelligent, and considerate language interactions. It can provide users with services in their learning, work, and life, including but not limited to Q&A, text generation, image generation, processing documents, and emotional chats, thus bringing users a more convenient and enjoyable intelligent life experience.

In terms of cooperation with ecosystem partners, vivo has independently developed matured large AI model capabilities on the device side. It plans to make these capabilities available to ecosystem partners in the near future. vivo has also open-sourced a 7B model and will make the 1+N architecture based on the 7B model available for internet partners to develop large AI model-based business applications.

To better serve global consumers, smartphone OEMs adhere to an open, cooperative, and win-win philosophy in adopting large AI models and developing the application ecosystem. They are actively collaborating with third-party large AI models and cloud service providers, providing necessary technical and financial support for developers engaged in generative AI application R&D.

Counterpoint Research expects leading smartphone OEMs to continue to optimize their own AI large models by either building their own cloud computing centers or renting third-party cloud computing resources. Additionally, the ability to form deep, strategic partnerships with industry ecosystem partners on AI large models and generative AI applications will also impact each OEM’s performance in the era of generative AI.

Generative AI Strategies of Developers

Developers will, undoubtedly, be significant contributors to the generative AI application ecosystem. Smartphones have penetrated into every aspect of people's lives and work. The broad application ecosystem has provided consumers services for socializing, entertainment, travel, mobile payments, online shopping and even mobile office. Generative AI technology will empower the developer community, assisting app developers in renovating existing applications and creating stunning new applications based on LLMs. This will lead to the birth of a new application ecosystem atop generative AI phones, making these smartphones a vital link between consumers and the digital world.

Huya

Huya Live is an interactive live-streaming platform with a focus on gaming. It has actively adopted generative AI technology in the live streaming field, with applications such as AI beauty, AI anchor assistants, AI audience assistants, personalized virtual anchors for individual viewers and an AI zone. All these are aimed at creating a personalized user experience unique to Huya for both broadcasters and viewers.

- AI Beauty: Using image generation capabilities, Huya has developed an AI beauty feature that can perform low-latency AI beautification on mobile devices, increasing the number of talent-focused vbloggers.

- AI Anchor Assistant: Based on multimodal large models, the AI assistant can understand the content and the game broadcasters' speech in real time. It interacts with the audience through live comments, explains game events, and initiates chat topics, enhancing the interactive experience during the streaming.

- AI Audience Assistant: The AI audience assistant can analyze viewers' watching history to understand their preferences and recommend interesting or useful live streaming via text and voice. This new interactive recommendation approach addresses the weakness of traditional algorithm which cannot accurately understand viewer intentions or cannot provide recommendations in an interactive style.

- Personalized Broadcasting Anchor: Based on multimodal large models, viewers can customize the anchor's personality, appearance and streaming content to generate a virtual AI anchor. Viewers can also interact in real-time with it. Additionally, generative AI phones equipped with LLMs locally can also modify the live anchor's appearance, voice and visual style according to viewer preferences, allowing a single anchor to provide a personalized viewing experience to different viewers.

- AI Zone: Huya has created an AI zone to enrich the audience's experience before and after watching live streams. The AI zone includes an AI camera feature based on the capability of text-to-image generation, LLM-based text challenge games, role-playing, and anchor avatars, among other features.

- Furthermore, Huya is highly focused on the development of AI agents and video generation. These advanced capabilities could fundamentally change the content production mode and interactive experience for the broadcasting industry, bringing about richer application scenarios.

Kugou

In the field of music, large AI models are capable of bringing in new value. As a Chinese leading digital music interaction service platform, Kugou Music is actively exploring how to empower music content creation using generative AI technology and making the music experience personalized. Leveraging the multimodal capabilities of large AI models, Kugou Music enables musicians to create more higher-quality music content, including personalized content creations. Music can also serve as an input so AI could

provide a comprehensive description and generate personalized playlists to help listeners better appreciate the charm of music. Using large AI models deployed in the device-cloud hybrid mode, users can enjoy music creation, classification, recommendation, and style transformation. This will lead to brand-new business models and accordant commercialization modes, thus bringing about an upgrade in the music industry.

Soul

Soul is a novel social platform launched in 2016. The AI-native social network is one of the few applications that provide real-time communication and interaction experiences with virtual personas. In the era of AI social networking, key aspects powered by AI such as relationship recommendations, conversation aids, lowering barriers to expression, and enhancing social experiences are critical for AI Native social networks and are significant opportunities for Soul.

In 2020, Soul initiated its R&D on AI-generated content (AIGC), gaining expertise in intelligent dialogue, image generation and voice technology. In 2023, Soul officially launched its self-developed LLM, SoulX. Addressing the core social needs of the younger generation, Soul has progressively introduced and internally tested several tools and innovative functions such as the intelligent dialogue robot "AI Goudan," AI-assisted chatting, and virtual companionship, further enriching the social experiences of its app users. Additionally, Soul is exploring the integration of more AI capabilities into its products, including developing companion AI to enhance gamified social experiences, tools for generating images and videos based on texts to lower expression barriers, and enhancing interactive and immerse experience with virtual personas. Simultaneously, Soul has been focusing on and following up on the development of AI technology on devices with inference capability, thus better enhancing user experience and providing privacy protection.

Tencent AI Lab

Since 2016, Tencent AI Lab has been dedicated to AI research for games and has released multiple world-leading scientific achievements such as Fine Art, Rlogist and Tencent AI Arena, impacting hundreds of millions of players. At the 2024 Game Developers Conference (GDC), Tencent AI Lab launched its self-developed game AI engine, GiiNEX, which is based on generative AI and decision AI technologies, providing a rich suite of AI solutions for the entire lifecycle of games. In terms of generative AI, GiiNEX can help in offering capabilities for AI NPCs, scene production and content generation. It also provides diverse generative AI capabilities including 2D images, animations, 3D cities, narratives, dialogues, game levels and music, enhancing developers' efficiency in producing high-quality content.

GiiNEX includes three cores – cutting-edge algorithmic models, an efficient training platform, and an online inference engine, supporting the entire lifecycle of a game from development to operation. The foundational AI research capabilities including reinforcement learning and natural language processing form a unified underlying algorithmic model that supports over ten types of games, such as MOBA, FPS, and party games. The high-performance training platform, custom-designed for agents and large models, supports resource scheduling on a scale of tens of thousands of computing boards. The online inference

engine, compatible with mainstream devices, achieves hybrid deployment on both mobile and cloud platforms, ensuring multi-device collaboration.

GiiNEX is committed to helping developers create the vibrant game world. For mobile gaming scenarios, GiiNEX has developed an intelligent tutoring feature that allows AI to serve as a personal coach for players. Moreover, the UGC game-level generation solution provides players with a suite of in-game AI tools, including text-to-image for inspiration, procedural content generation (PCG) of buildings, 3D model assembly and generation, color scheme generation, and NPC action generation, helping players enhance their creative efficiency and enrich personalized content.

Currently, the development of large AI models is advancing rapidly and the number of models is also increasing fast. However, deploying and applying these models on mobile devices is still at an early exploratory stage. Application developers, including Tencent, Huya, Kugou and Soul, are closely collaborating with smartphone OEMs and chip vendors. On one hand, they are developing upper-layer applications based on LLMs pre-installed on mobile phones, allowing users to truly experience the allure of generative AI. On the other hand, developers can deeply participate in model development, creating large AI models tailored to specific business scenarios, collectively striving to build the ecosystem for generative AI phones.

Chapter Three: Hardware and Software Technology Landscape of Generative AI Phones

SoCs/APs are the most critical AI computing units in generative AI phones, largely determining the parameter size of large AI models deployed on smartphones. They, in turn, somehow determines the maximum capabilities a generative AI phone can support. Besides hardware support, complete toolkit support is also indispensable. This includes software development kits (SDKs) and pre-trained large models to assist developers in rapidly developing applications at various use scenarios.

This section will elaborate on the requirements of generative AI in terms of smartphone hardware specifications, as well as the importance of the accompanying software ecosystem.

Hardware Requirements for Deploying Large AI Models

The capabilities and performance of large AI models are largely determined by the quantity and quality of parameters. Generally, the more parameters a model has, the more accurate and the more powerful the model is. The number of parameters also determines the smartness of the generative AI phone.

However, the smartphone hardware design is quite demanding. It needs to take into consideration constraints such as heat dissipation, power consumption and battery life, as well as the PCB footprint. Additionally, to enable consumers to have a distinctive experience of generative AI capabilities, the parameter size of the LLMs running on the smartphone cannot be too low, and it needs to support multimodal and multilingual features. Furthermore, consumers generally expect the response speed of on-device large models to be almost "real-time". Taking MediaTek's Dimensity 9300 as an example, inference speed for a 7-billion-parameter on-device LLM can achieve up to 20 tokens per second, well matching the average reading speed of humans.

The AI computing power of SoC platforms has become an increasingly important performance metric for chip vendors, smartphone OEMs and consumers. Many smartphone SoCs have integrated independent AI computing units like APU or NPU, dedicated to handling heavy AI tasks. At the same time, chip vendors are redesigning these AI processing units to better support the Transformer architecture, thus achieving higher trillion operations per second (TOPs) per watt within a 1-watt to 2-watt power budget. This is crucial for deploying and running on-device large AI models and also in helping smartphones handle more complex generative AI tasks.

Currently, smartphone SoCs integrate heterogeneous processing units, including CPUs, GPUs, DSPs, and dedicated AI computing units. Counterpoint Research believes that generative AI phones in the future will need to run multiple AI models simultaneously, including large multimodal models as well as smaller models for photography enhancement and image generation. The system will call upon the most suitable unit to process the tasks, thus achieving lower power consumption and latency. So, dedicated AI computing units and heterogeneous computing architectures will coexist in smartphone SoCs, and the importance of sufficient communication bandwidth between different computing units will also rise.

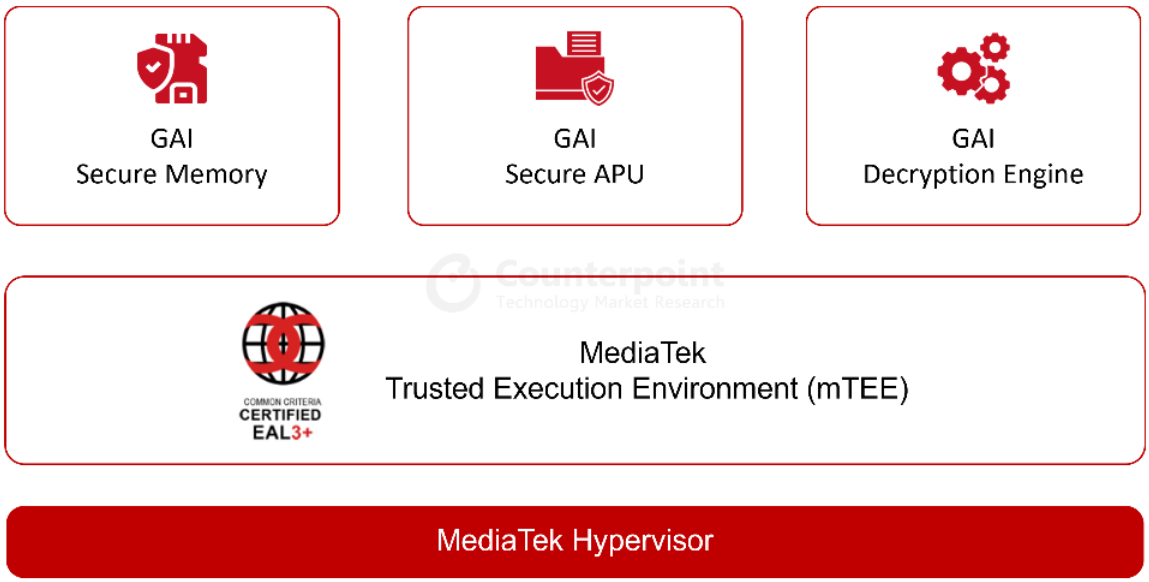
Moreover, a large AI model running on device resides in the memory, so each generative AI task may involve massive data transfer. Accordingly, generative AI phones demand high memory capacity and bandwidth. Taking the current mainstream model with 7 billion parameters as an example, about 4GB of DRAM space is required to run the model, and it is recommended to use at least 8GB of LPDDR5x (I/O bandwidth is recommended to be over 60GB/s). Thus, a high-performance SoC platform and high-specification memory are both indispensable for generative AI phones. Therefore, the ideal user experience can be achieved only with the proper combination.

Also, compared to large AI models on the cloud, large AI models on device store and process user privacy data entirely on the device, which makes it much more secure. However, chip vendors should also provide security solutions that will ensure users the benefits of generative AI while also keeping data and privacy safe.

In response to generative AI, MediaTek has upgraded its existing mobile hardware-level security mechanism to generative AI-level security solutions, providing comprehensive protection from model parameters to user privacy, and operational data. Given the massive data storage requirements for generative AI, MediaTek has improved its AI security architecture by upgrading its Secure Memory, Secure APU, and Secure Decryption Engine, which has stepped up the upper limit of storage capacity, the efficiency of APU operations, and decryption speed. The new AI Security 2.0 represents a significant upgrade and takes generative AI-level security to new heights.

Exhibit 5 MediaTek's Generative AI Security Solution

Target (AI Security version)		AI Security 1.0	AI Security 2.0
mTEE cc EAL 3+Certificated	Secure memory allocation	V4.0 Max. 3.8 GB (0.38s for 1GB allocation)	V5.0 Max. 30 GB (0.05s for 1GB allocation)
Decryption	Gigabytes Data	**0.3 GB/s** (Big core CPU)	**3 GB/s** (CPU crypto ext.)
Secure APU	AI Performance	**15% perf. Drop up to 1 VM**	**5% perf. Drop up to 1 VM**



Sources: MediaTek, Counterpoint Research

Finally, accessing large AI models on the cloud requires smartphones to feature a fast and stable network connection. 5G network allows smartphones to upload and download data more quickly, which is particularly important for AI applications that rely on cloud-based large AI models. Users can access cloud-based AI models more rapidly, and the results can be returned to users within reasonable latency.

Demand for Software Ecosystem

Deploying large AI models on smartphones needs more than just support from the hardware. Support from the software ecosystem is also essential. The ecosystem helps AI models better adapt to the actual hardware conditions of current smartphones and provides developers and OEMs with an integrated development environment. It aids in the rapid development and deployment of generative AI applications on mobile devices, promoting the continued healthy and positive development of the industry.

First, smartphone memory storage, especially RAM space, is limited. It is necessary to use techniques like Quantization and Pruning to optimize the size of pre-trained AI models so that they take up less memory space. Quantization is a model compression technique aimed at reducing model size and accelerating inference speed while maintaining the accuracy of model outputs. AI large models typically use 32-bit floating-point data (FP32) type. The mainstream solutions now tend to convert them to INT4 or INT8 integer type. For example, a 13 billion parameter large AI model can be compressed from a size of tens of GB to less than 13GB after quantization. However, when considering the use of the INT4 quantization algorithm to compress models, it is necessary to ensure that the SoC platform supports such low-precision operations. Moreover, chip vendors enhance the execution speed on existing hardware by adopting fast decoding technologies, including speculative decoding, Medusa, and more.

Moreover, for application developers, if chip vendors can provide pre-trained generative AI large models and fine-tune them for the SoC platform, the application development process can be shortened significantly and the resulting faster AI inference speed can offer a better user experience.

Currently, large generative AI models, especially LLMs, are emerging rapidly. Many companies from China and abroad, including Google, Meta, Alibaba and Baichuan have launched generative large AI models for smartphone deployment. Those companies have worked with chip vendors to optimize the models and help partners to fast deploy models quickly.

Privacy has been an increasing concern for consumers, so developers need to incorporate privacy protection into their overall product design. If chip vendors can provide developers with system-level security solutions, it could help developers shorten the application development cycle, ease consumer privacy concerns and achieve a win-win situation for developers and chipset vendors.

Given the high complexity of generative AI technologies, collaboration in all sectors is required to propel the stable and rapid development of the industry. Particularly in the field of large AI models, the high threshold for training large models not only creates opportunities for developers of large models but also increases the complexity of collaboration among chip vendors, smartphone OEMs, large AI model providers, and application developers. To better facilitate the cooperation of multiple parties, MediaTek has initiated the Dimensity AI Pioneer Project and invited leading Chinese smartphone OEMs, large AI model developers, and innovative application developers to join it. This initiative aims to accelerate the

application of generative AI technologies on smartphones, thereby providing consumers with more AI features and a better user experience.

Current Smartphone SoC Platforms Support AI on Device

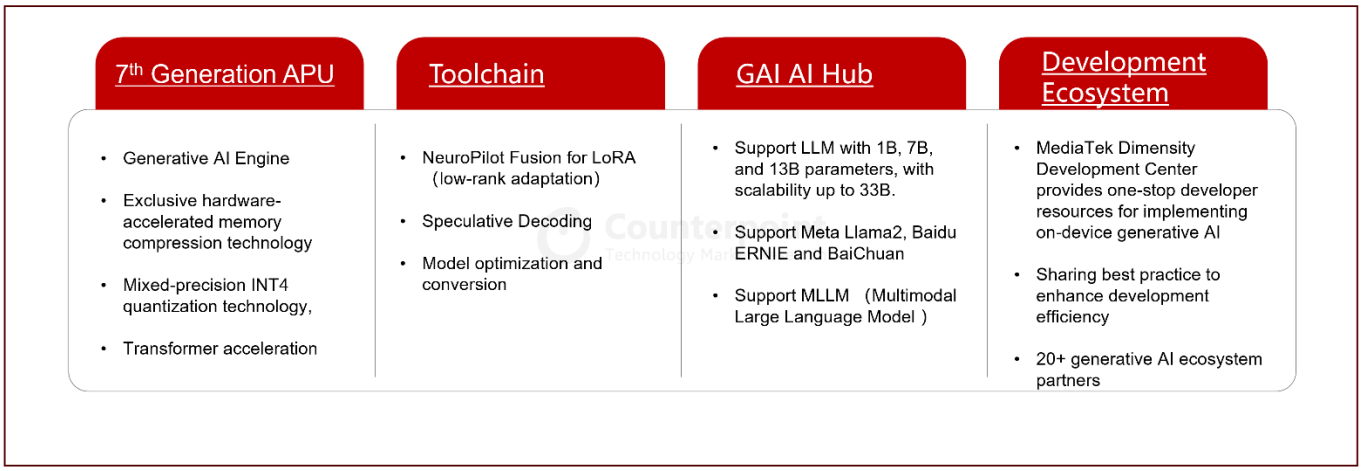
The smartphone industry's adoption of generative AI technology is still at an early stage. The deployment of generative AI models on smartphones has introduced new requirements for smartphone specifications, which is a costly affair. Therefore, integrating on-device large AI models with smartphones is seen mostly in high-end product lines. As leading chip vendors, MediaTek and Qualcomm have both provided powerful mobile computing platforms that support deploying multimodal large AI models on device. They also have fine-tuned current on-device large AI models.

In November 2023, MediaTek released the Dimensity 9300 flagship mobile SoC platform, integrating MediaTek's seventh-generation AI processor called the APU 790. According to MediaTek, the APU 790 doubled the integer and floating-point computing power compared to its predecessor while reducing power consumption by 45%. Moreover, the APU 790 includes a hardware-level generative AI engine, enabling faster and more secure edge AI computing. Compared to the previous generation, it specifically accelerates operations for the Transformer architecture which is commonly used in current LLMs. The processing speed for large models is eight times faster than its predecessor.

The Dimensity 9300 was also among the first SoC to support LPDDR5T technology, with a transmission rate up to 9600Mbps, allowing rapid data transfer between memory and processor. This improves the response speed of on-device generative AI models and allows devices to run complex AI algorithms and models more effectively. Despite the faster transmission rate, LPDDR5T technology still retains low power consumption characteristics.

Beyond hardware, MediaTek is committed to building a favorable development environment for developers. MediaTek has launched its generative AI solution NeuroPilot for developers. NeuroPilot is MediaTek's core for building a generative AI ecosystem. It currently includes MediaTek's latest seventh-generation APU, toolkits, a developing model center (GAI Model Hub), and the development ecosystem.

Exhibit 6 MediaTek's Ecosystem for AI Development



Sources: MediaTek, Counterpoint Research

The Neuropilot Compression, a hardware-level compression technology, can further compress the model size based on quantization, alleviating the pressure on memory bandwidth when running large AI models on device. NeuroPilot also utilizes Low-Rank Adaptation (LoRA) technology to help large AI models on device expand their capabilities without redeploying multiple models, significantly reducing development complexity and the demand for smartphone storage space. MediaTek has also established the GAI Model Hub for developers, containing a rich set of AI models for specific problems, enabling developers to use these AI models conveniently and quickly. With the ecosystem provided by Neuropilot, developers can quickly and efficiently deploy generative AI on device.

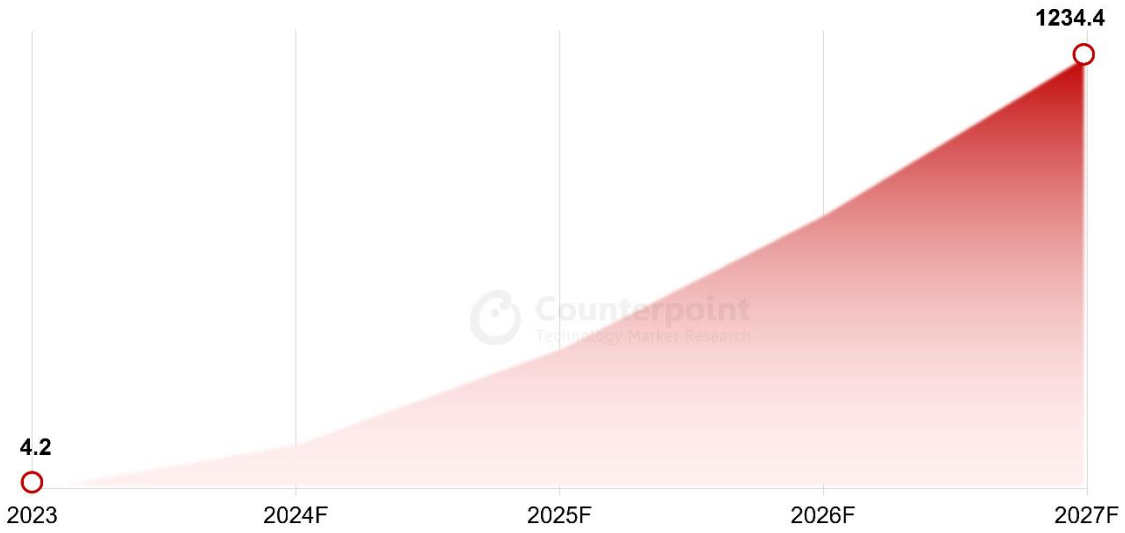
Qualcomm released the Snapdragon 8 Gen 3 in 2023 as well. The firm also optimized the design for generative AI. According to Qualcomm, the NPU performance improved by 98% compared to the previous generation, and the performance per watt increased by 40% for continuous AI inference scenarios. The Snapdragon 8 Gen 3 also supports LPDDR5x, with a data transmission rate of up to 8500Mbps, making it more suitable for generative AI on device than its predecessor. In terms of the developer ecosystem, Qualcomm launched the Qualcomm AI Hub, which currently includes 75 AI models that can be deployed on Qualcomm-supported devices.

Chapter Four: Generative AI Phone Forecast

The integration of generative AI with smartphones initially started with flagship product lines. According to Counterpoint Research, of the 1.17 billion smartphones shipped in 2023, less than 1% met Counterpoint's definition of a generative AI phone. However, by 2024, thanks to the supply chain actively embracing the technology revolution, major OEMs are prioritizing generative AI capabilities when upgrading the mid-to-high-end and high-end products. This is expected to accelerate the proliferation of generative AI phones, and the penetration rate of shipment is expected to reach 43% by 2027. At the same time, Counterpoint Research predicts that the installed base of generative AI phones will grow from millions of units in 2023 to about 1.23 billion units by 2027.

Exhibit 7 Generative AI Phone Installed Base Forecast

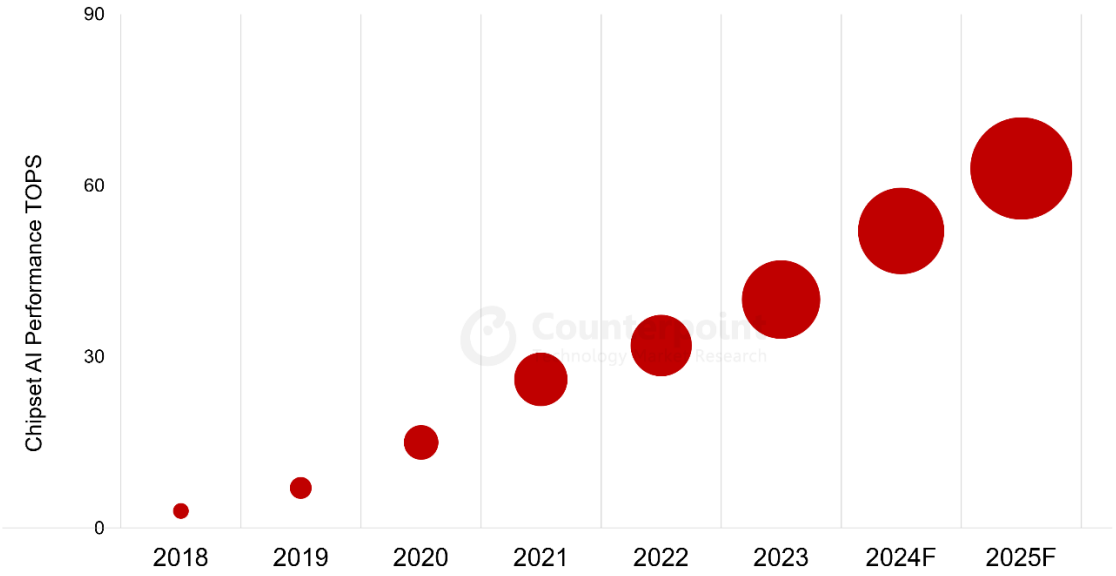
In Million Units



Source: Counterpoint Research AI 360 Service

The AI capabilities of generative AI phones are closely related to the TOPS performance of an SoC. The AI computing power of flagship smartphones, measured in TOPS, has increased 20 times since 2017. In this context the AI capabilities of smartphones are becoming increasingly powerful. Chip vendors have played a significant role in this transformation. Counterpoint Research predicts that SoCs' AI computing limits for flagship smartphones will continue to grow to more than 60 TOPS by 2025.

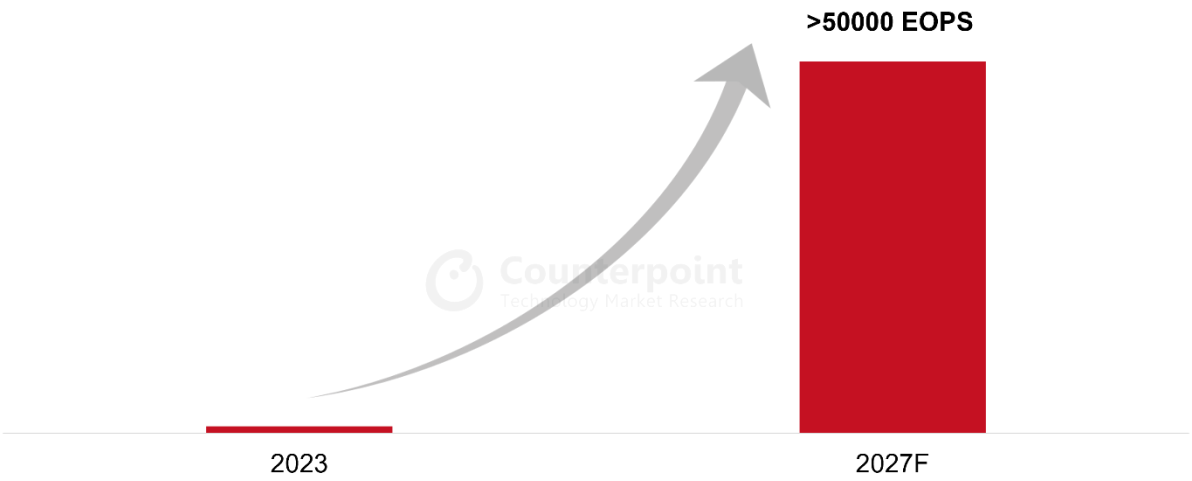
Exhibit 8 Smartphone Peak AI Computing Power (by TOPS)



Source: Counterpoint Research AI 360 Service

As mentioned earlier, the installed base of generative AI phones is continuously growing, and the AI computing power per device is also increasing. Combined, these factors will lead to explosive growth in the overall AI computing power of global smartphones. According to Counterpoint Research's estimates, by 2027, the global AI computing pool of generative AI phones will exceed 50,000 EOPS. In the future, generative AI phones will become an integral part of global AI computing resources, further confirming the feasibility, cost-effectiveness, and necessity of deploying AI in the device-cloud hybrid mode.

Exhibit 9 Total AI Computing Resources on Generative AI phones



Source: Counterpoint AI 360 Service

Conclusion

Generative AI phones have initiated a new cycle in the intelligence evolution of smartphones. With the joint efforts of chip vendors, smartphone OEMs, large AI model developers, and numerous developers, generative AI phones will provide users with brand-new interactive experiences, multimodal content generation capabilities, personalized services, and a revolutionary application ecosystem. Benefiting from ever-increasing AI computing power on device and the growing capabilities of large AI models, the integration of generative AI technology with smartphones will deepen further. Generative AI phones will evolve into mobile intelligent agents, opening up new imaginations. This will support and sustain the prosperity of the smartphone industry and the mobile application ecosystem, and help the industry better prepare for the era of 6G.

Authors, Copyright, User Agreement and Other General Information

Tarun Pathak
Research Director
tarun@counterpointresearch.com

Ethan Qi
Associate Director
ethan@counterpointresearch.com

Archie Zhang
Research Analyst
archie.zhang@counterpointresearch.com

COUNTERPOINT TECHNOLOGY MARKET RESEARCH
Hong Kong | USA | South Korea | India | UK | Argentina | China
info@counterpointresearch.com