Brought to you by Informa Tech

GenAI in the Smartphone

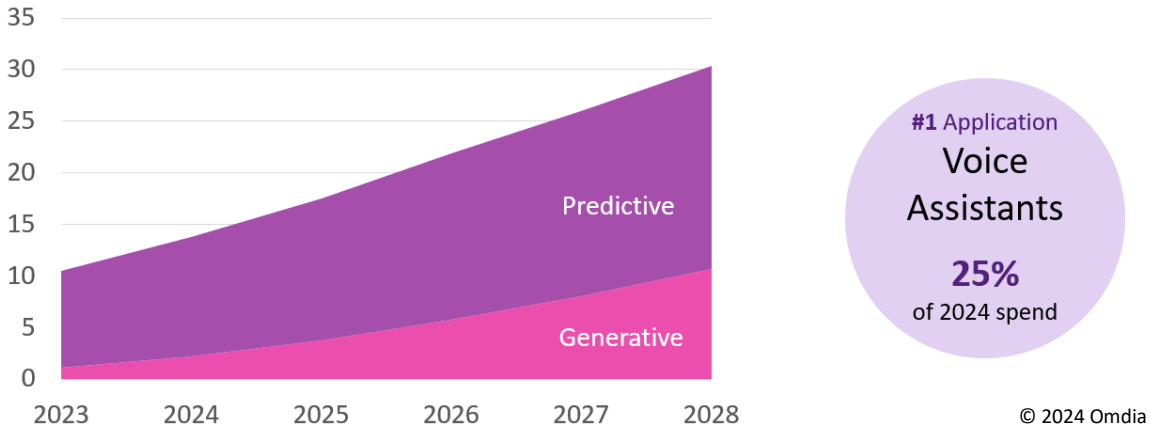
Publication date:
June 2024
Author(s):
Alexander Harrowell, Principal Analyst, Advanced Computing for AI
Josh Builta, Senior Research Director, AI & IoT

The rapid rise of GenAI in smartphones

The era of generative AI (GenAI) has already changed the way consumers interact with their favorite brands, upended the way companies think about corporate data, and even changed the concept of work itself—never before in the technology market has a single idea so quickly captured the attention of buyers and builders alike. Omdia research estimates an overall market valuation going from zero in late 2022 to $74bn by 2028.

Top among the industries served by AI is the consumer (digital) industry, rising from a $7.7bn spend in 2022 to $30.4bn in 2028, with smartphones behind the majority share of the deployment. The top use cases will include voice assistants, search engines, writing assistants, and voice/speech recognition, as well as image recognition, classification, and description.

Figure 1: Annual growth rate for AI software spend in the consumer (digital) industry



Source: Omdia AI Software Market Forecast. Consumer (Digital) is defined as specifically focused on the use of AI in products or services that are directly marketed at and used by consumers with a focus on internet services, as opposed to other enterprise or business customers. AI technology is often embedded in devices, such as smartphones, home assistants, or content services, and uses personal preference and behavioral data to create a more intelligent, personalized experience.

The hardware view

While consumers' insatiable appetite for GenAI applications grows, the focus shifts for device manufacturers to the "how"—i.e., AI processors in the smartphone.

Although AI accelerator adoption in smartphones has been faster than is widely realized, weak demand for smartphones, in general, has slowed down adoption in the last two years, sticking around 65% (attach rate of AI accelerators to smartphones). Despite this setback, AI accelerators are still making their way into cheaper and cheaper devices, while products introduced in 2018–20 are now aging out and are being replaced by considerably more powerful system-on-chips (SoCs).

Omdia expects the combination of smartphones moving upmarket and accelerators downmarket will eventually push adoption up over 90%.

Omdia expects the combination of smartphones moving upmarket and accelerators downmarket will eventually push adoption up over 90%. The bulk of the growth, though, is in the top performance classes—30–50 and 50+ TOPS, or in other words, the devices capable of running smaller GenAI models. This is in part due to the emergence of GenAI, in part due to the development of stronger SoCs such as Qualcomm's Snapdragon 8 Gen3, and in part due to so-called "premiumization" of the smartphone market, which is growing more strongly in the higher price bands.

In particular, Omdia sees three key drivers behind the upward trend:

- **Adoption:** Image-centric applications and voice-based assistants are crucial for a smartphone's competitive position. As such, acceleration is valuable for use cases such as super-resolution, photo editing, image/video search, and speech recognition. The growing class of "small LLMs" is becoming increasingly interesting for generative, multi-modal, and spatial use cases, with vendors targeting models in the 1B–10B range.

- **Composable SoC architectures:** The unique constraints of the mobile ecosystem mean that devices have always had to be SoCs. The major vendors in the Arm ecosystem, such as Qualcomm, Samsung, Apple, and MediaTek, have significant expertise in integrating heterogeneous computing cores into SoCs, including trusted enclaves, MEMS, radios, GPUs, cameras, and AI accelerators. Although AI has been integrated into sensor, radio, camera, and audio functions, scaling logic now encourages vendors to pull these into a bigger central AI accelerator block.

- **Pursuit of differentiation:** Convergent smartphone design has led vendors to seek differentiation through hard-to-replicate deep technology.

The trend toward smaller models is particularly significant—Omdia sees the future of AI models as specialized, multi-modal, and fine-tuned on the user's own data. As only the most premium phones can have powerful multi-modal AI models on-device, many other (and older) phones will have to rely on a lighter AI engine for non-generative tasks and instead use cloud computing for more intensive tasks. As this will be lighter for the processor, it will be better for more efficient battery consumption. Even Apple, a premium vendor if ever there was one and having had the Apple Neural Engine ever since 2017's A11 Bionic, has chosen to offload some of the tasks in Apple Intelligence to its new private cloud service.

Ultimately, hardware will reflect the trend of smaller large language models (LLMs) and a combination of both cloud and edge deployments. A revolution in open-source AI that shifted the focus of innovation

from building giant systems at companies like OpenAI or Google to building much smaller, often domain-specific models in the 5B–50B parameter range happened in 2023. This implies that both applications that might have needed a gigantic model can be served with something small enough to run on a PC or a smartphone, and that some applications that have been served with a relatively simple computer vision model (e.g., YOLO or ResNet) will be moving to so-called "small" LLMs to benefit from multi-modal capability, transfer learning, and plain-language prompt development. Simultaneously, developers are especially keen to run LLMs locally in inference and, to a lesser extent, for fine-tuning.

So where next?

Developers' experience is crucial. The four P's of AI at the edge are privacy, personalization, performance, and power efficiency, but a fifth P—developer productivity—is crucial to delivering them.

Developers' experience is crucial. The four P's of AI at the edge are privacy, personalization, performance, and power efficiency, but a fifth P—developer productivity—is crucial to delivering them.

There is a need for software tools that bridge the gap between cloud training and edge inference. There will be an architecture split between the data center and the edge—where the edge market exhibits nothing like the NVIDIA-led GPU hegemony in the data center. AI training generally happens in the data center, so this architecture split poses problems for software development. Developers will be even more important than usual, as they are completely open to what applications will use the increasing inference power available at the edge. Consequently, they will sell the hardware.

And with more tools (both hardware and software) to play with on the device, AI features will become more accessible to users of smartphones from top of the range down to mid-end devices in 2024 and beyond. When Samsung, Google, and Huawei were the first to introduce GenAI features, it was seen as a premium-only feature, but with the Pixel 8a, Samsung S22, and other devices now in the mid-end price tier getting GenAI features, this expansion is happening faster than many anticipated. Hybrid AI will be a key part of this—where older or mid-range chipsets cannot effectively run LLMs, more phones will rely on lighter AI models for on-device features and cloud AI computing for bigger GenAI tasks.

The snowball effect, therefore, continues—especially as third-party developers join in to help create new AI experiences for consumers, alongside and in addition to the wave of OEM introductions—just one of the reasons why Omdia says the consumer industry will remain the leader of the AI pack through 2028 and beyond.

Authors

Alexander Harrowell, Principal Analyst, Advanced Computing for AI
Alexander.Harrowell@omdia.com

Josh Builta, Senior Research Director, AI & IoT
Joshua.Builta@Omdia.com

This piece of research was commissioned by Qualcomm Technologies, Inc.

Request external citation and usage of Omdia research and data via citations@omdia.com.

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help you. For more information about Omdia's consulting capabilities, please contact us directly at consulting@omdia.com.

CONTACT US
omdia.com

customersuccess@omdia.com