
Qualcomm
May 2023
The future of AI is hybrid
Part I:
Unlocking the generative AI future with on-device and hybrid AI

Table of contents

1 Executive summary

The future of AI is hybrid. As generative AI adoption grows at record-setting speeds[1] and drives higher demand for compute,[2] AI processing must be distributed between the cloud and devices for AI to scale and reach its full potential – just like traditional computing evolved from mainframes and thin clients to today's mix of cloud and edge devices. A hybrid AI architecture distributes and coordinates AI workloads among cloud and edge devices, rather than processing in the cloud alone. The cloud and edge devices, such as smartphones, vehicles, PCs, and IoT devices, work together to deliver more powerful, efficient, and highly optimized AI.

The main motivation is cost savings. For instance, generative AI-based search cost per query is estimated to increase by 10 times compared to traditional search methods[3] – and this is just one of many generative AI applications. Hybrid AI will allow generative AI developers and providers to take advantage of the compute capabilities available in edge devices to reduce costs. A hybrid AI architecture (or running AI on device alone) offers the additional benefits of performance, personalization, privacy, and security – at a global scale.

Hybrid AI architectures can have different offload options to distribute processing among cloud and devices depending on factors such as model and query complexity. For example, if the model size, prompt, and generation length is less than a certain threshold and provides acceptable accuracy, inference can run completely on the device. If the task is more complex, the model can run across cloud and devices. Hybrid AI even allows for devices and cloud to run models concurrently – with devices running 'light' versions of the model while the cloud processes multiple tokens of the 'full' model in parallel and corrects the device answers if needed.

The potential of hybrid AI grows further as powerful generative AI models become smaller while on-device processing capabilities continue to improve. AI models with over 1 billion parameters are already running on phones with performance and accuracy levels similar to those of the cloud, and models with 10 billion parameters or more are slated to run on devices in the near future.

The hybrid AI approach is applicable to virtually all generative AI applications and device segments – including phones, laptops, XR headsets, vehicles, and IoT. The approach is crucial for generative AI to scale and meet enterprise and consumer needs globally.

[1] https://www.statista.com/chart/29174/time-to-one-million-users/
[2] https://siliconangle.com/2023/02/05/generative-ai-drives-explosion-compute-looming-need-sustainable-ai/
[3] https://www.reuters.com/technology/tech-giants-ai-like-bing-bard-poses-billion-dollar-search-problem-2023-02-22/

2 Introduction and current trends for generative AI

ChatGPT has captured our imagination and engaged our curiosity. Reaching 100 million monthly active users just two months after it launched in November 2022, it is the fastest growing consumer app in history – and the first generative AI "killer" app. With the rapid pace of innovation, it is becoming more and more difficult to keep up with generative AI developments. According to a major aggregator site, there are over 3,000 generative AI apps and features available.[4] AI is having a Big Bang moment, akin to the launch of television, the worldwide web, or the smartphone. And this is just the beginning.

Generative AI models, such as ChatGPT and Stable Diffusion, can create new and original content like text, images, video, audio, or other data from simple prompts. These models are disrupting traditional methods of search, content creation, and recommendation systems – offering significant enhancements in utility, productivity, and entertainment with use cases across industries, from the commonplace to the creative. Architects and artists can explore new ideas, while engineers can create code more efficiently. Virtually any field that works with words, images, video, and automation can benefit.

Web search is one of many applications being disrupted by generative AI. Another example is Microsoft 365 Copilot, a new productivity feature that uses generative AI to help write and summarize documents, analyze data, or turn simple written ideas into presentations – all embedded in Microsoft apps including Word, Excel, PowerPoint, Outlook, Teams, and more.

The emergence of generative AI also marks the first step toward a more diverse and personalized digital landscape for users to explore. It has the potential to democratize 3D content creation since 3D designers can develop 3D content faster and more efficiently with generative AI tools. This will not only accelerate the creation of immersive virtual experiences, but also reduce the barriers to entry for individual creators to produce their own content.

We are on the cusp of seeing a variety of new enterprise and consumer use cases emerge from generative AI, delivering capabilities we have yet to even imagine. Foundation models, such as general-purpose large language models (LLMs) like GPT-4 and LaMDA, have achieved unprecedented levels of language understanding, generation capabilities, and world knowledge. Most of these models are quite large, with more than 100 billion parameters, and expose their functionality to customers via APIs, free or paid.

Access to foundation models is leading to a deluge of both startups and large organizations building apps across text, image, video, 3D, speech, and audio. Examples include code generation (GitHub Copilot), text generation (Jasper), image generation for artists and designers (Midjourney), and conversational chatbots (Character.ai).

[4] Generative AI apps and features as of April 2023 on https://theresanaiforthat.com/

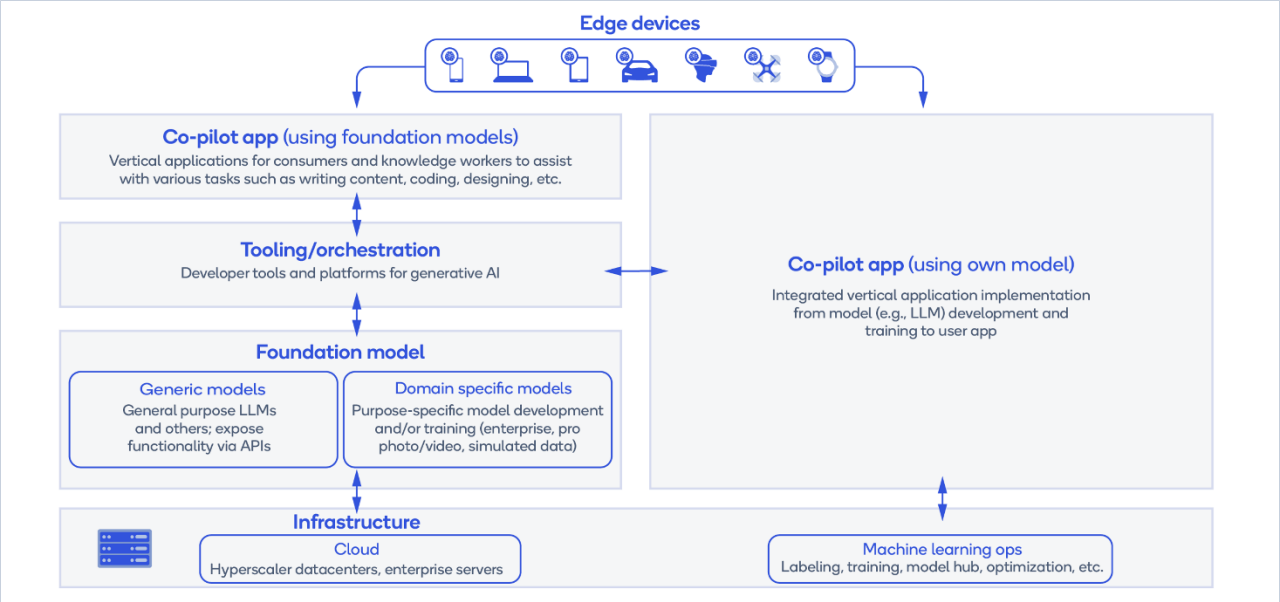


Figure 1: The generative AI ecosystem stack is allowing many apps to proliferate.

Early estimates value the market size of generative AI at $1 trillion[5] – spread through participants in the ecosystem stack. To realize this massive opportunity and drive AI into the mainstream, the computing architecture needs to evolve and meet the increased processing and performance demands of generative AI at scale.

3 Hybrid AI is crucial for generative AI to scale

Massive generative AI models with billions of parameters place significant demands on computing infrastructure. As such, both AI training, which learns the parameters for an AI model, and AI inference, which executes the model, have been constrained to cloud implementations for large and complex models – until now.

The scale of AI inference is poised to be significantly higher than that of AI training. Although training individual models consumes significant resources, large generative AI models are expected to be trained only a few times a year. However, the cost of inferencing with those models increases correspondingly with the number of daily active users and their frequency of use. Running inference in the cloud results in very high costs that can be unsustainable for scaling.

Hybrid AI is the solution – just like traditional computing went from mainframes and thin clients to a mix of cloud infrastructure and smart devices including PCs and smartphones.

3.1 What is hybrid AI?

Hybrid AI is the device and cloud working together, splitting AI computation where and when appropriate, to provide enhanced experiences and efficient use of resources. For some scenarios, the computation will be centered primarily around the device, offloading tasks to the

[5] UBS, Feb 2023

cloud when necessary. In cloud-centric implementations, devices will opportunistically offload AI workloads from the cloud, where possible and subject to their capabilities.

3.2 Benefits of hybrid AI

A hybrid AI architecture (or running AI on-device alone) offers benefits with regards to cost, energy, performance, privacy, security, and personalization – at a global scale.

3.2.1 Cost

With generative AI model usage and complexity continuing to grow, running inference exclusively in the cloud is not economical due to increasing costs for data center infrastructure, including costs for hardware, real estate, energy, operations, additional bandwidth, and network ingress/egress fees.

For example, the current cloud-based computing architecture for LLM inferencing leads to higher operating costs for internet search companies, big and small. Consider a future with internet search augmented by generative AI LLMs, like GPT running with well over 175 billion parameters. Generative AI searches can provide a much better user experience and results, but the cost per query is estimated to increase by 10 times compared to traditional search methods, if not more. With more than 10 billion search queries per day currently, even if LLM-based searches take just a small fraction of queries, the incremental cost could be multiple billions of dollars annually.[6]

Shifting some processing from the cloud to edge devices reduces strain in cloud infrastructure and mitigates expenses. This makes hybrid AI crucial for generative AI to continue to scale – taking advantage of the billions of edge devices with AI capabilities already deployed, as well as the billions more coming with increased processing power.

Cost savings is also a critical enabler to the generative AI ecosystem, allowing OEMs, independent software developers (ISVs), and app developers to experiment and create apps more economically. For example, a developer could create an app based on Stable Diffusion that runs completely on the device and pay a lower per-query cost, or no cost at all, for every image generated.

3.2.2 Energy

Edge devices with efficient AI processing offer leading performance per watt, especially when compared with the cloud. Edge devices can run generative AI models at a fraction of the energy, especially when considering not only processing but also data transport. This difference is significant in energy costs as well as helping cloud providers offload datacenter energy consumption to meet their environmental and sustainability goals.

3.2.3 Reliability, performance, and latency

In a hybrid AI architecture, on-device AI processing provides reliable performance that can be comparable to the cloud or even better when cloud servers and network connectivity are congested. [7] Peaks in cloud demand for generative AI queries can create large queues and high

[6] Morgan Stanley, "How Large are the Incremental AI Costs...and 4 Factors to Watch Next" Feb 2023
[7] https://www.qualcomm.com/news/onq/2023/02/worlds-first-on-device-demonstration-of-stable-diffusion-on-android

latency, and in some cases even lead to denial of service.[8] This can be prevented by shifting compute loads toward edge devices. Additionally, the availability of on-device processing in a hybrid AI architecture can allow generative AI apps to work anywhere users are, even without connectivity.

3.2.4 Privacy and security

On-device AI inherently helps protect users' privacy since queries and personal information remain solely on the device. For enterprise and workplace usage of generative AI, this helps to solve the challenge of protecting company confidential information. For example, a programming assistant app for generating code could run on the device without exposing confidential information to the cloud – addressing a concern already faced by companies today.[9] For consumer usage, a "private mode" in a hybrid AI architecture allows users to strictly utilize on-device AI to enter sensitive prompts to chatbots, such as health questions or startup ideas. In addition, on-device security is strong and will evolve to ensure that personal data and model parameters are secure on edge devices.

3.2.5 Personalization

Hybrid AI makes more personalized experiences possible. Digital assistants will be customized to a user's expressions, quirks, and uniqueness without sacrificing privacy. This persona, which is a representation of a person based on real-life behaviors, values, pain points, needs, concerns, and problems, is learned and evolves over time. It is used to enhance and customize generative AI prompts, which are then processed either on-device or in the cloud. Since it stays within the device, and with on-device learning, the persona gets continuously optimized and updated.

Personalization is not just for consumers. For instance, organizations can use it to standardize how their code is written or to produce public content in a unique tone and voice.

3.3 How AI workloads are distributed

We envision a hybrid AI architecture with different offload options to distribute processing, which will evolve over time, depending on model and query complexity. For example, if the model size, prompt, generation length is less than a certain threshold and provides acceptable accuracy, then it can run completely on the device. If more complex, then the model can run jointly in the cloud and on device, or if more up-to-date information is required, then the internet can provide that information as well.

3.3.1 Device-centric hybrid AI

In a device-centric hybrid AI architecture, the device acts as the anchor point with the cloud used only to offload tasks that the device cannot sufficiently perform. Since many generative AI models can run adequately on the device (see Figure 2), the device can do most of the processing by running the less complex inferences.

For example, a user is running Microsoft 365 Copilot or Bing Chat on a laptop. Models with up to tens of billions of parameters will run on device while more complex models will use the

[8] https://www.digitaltrends.com/computing/chatgpt-is-at-capacity-and-is-frustrating-new-people-everywhere/

[9] https://www.pcmag.com/news/samsung-software-engineers-busted-for-pasting-proprietary-code-into-chatgpt

cloud as needed. The experience will be seamless to the user since an on-device neural network or rules-based arbiter will decide if the cloud is needed, whether for a better model or retrieving internet information. If the user is not satisfied with the quality of the results, a request to try again could initiate a better model to be used. As on-device AI processing improves with newer devices and more advanced chipsets, more can be offloaded from the cloud.

Figure 2: In a device-centric hybrid AI architecture, the cloud is only used to offload AI tasks that the device cannot sufficiently perform.

For various generative AI applications, such as creating images or draft emails, fast and responsive inferences are preferrable, even if that carries a small tradeoff with regards to accuracy. Quick feedback (i.e., low latency) with on-device AI allows users to quickly iterate with refined prompts until the output is satisfactory.

3.3.2 Device-sensing hybrid AI

In a device-sensing hybrid AI scenario, the edge models act as the sensor input – or the eyes and ears – to the cloud LLMs, which act as the brain. For example, a user speaks to their smartphone. An automatic speech recognition (ASR) AI model, such as Whisper, runs on device converting speech to text before sending it as a query to the cloud. The cloud runs an LLM and then sends the text answer back to the device. The device runs a text-to-speech (TTS) model to provide a natural and hands-free response. Offloading the ASR and TTS models to the device saves compute and connectivity bandwidth. As LLMs become multi-modal and allow images as input, the computer vision processing can also run on device, further offloading compute and reducing connectivity bandwidth – and therefore saving costs.

In a more advanced version that preserves privacy further, on-device AI takes on more processing and provides an improved and more personalized prompt to the cloud. Through on-device learning and personal data on the device, such as social media, email, messaging, calendar, and location, the device would create an individual persona of the user that works with an orchestrator program to provide improved prompts based on this additional context. For example, if a user asks their phone to schedule a time and reserve a table with their best friend at their favorite restaurant, the orchestrator knows those personal details and can provide a better prompt to the cloud LLM. The orchestrator can put guard rails on the LLM

when it lacks information and help prevent hallucinations. For simpler queries, a smaller LLM can run on device without any cloud interaction, similar to device-centric hybrid AI.

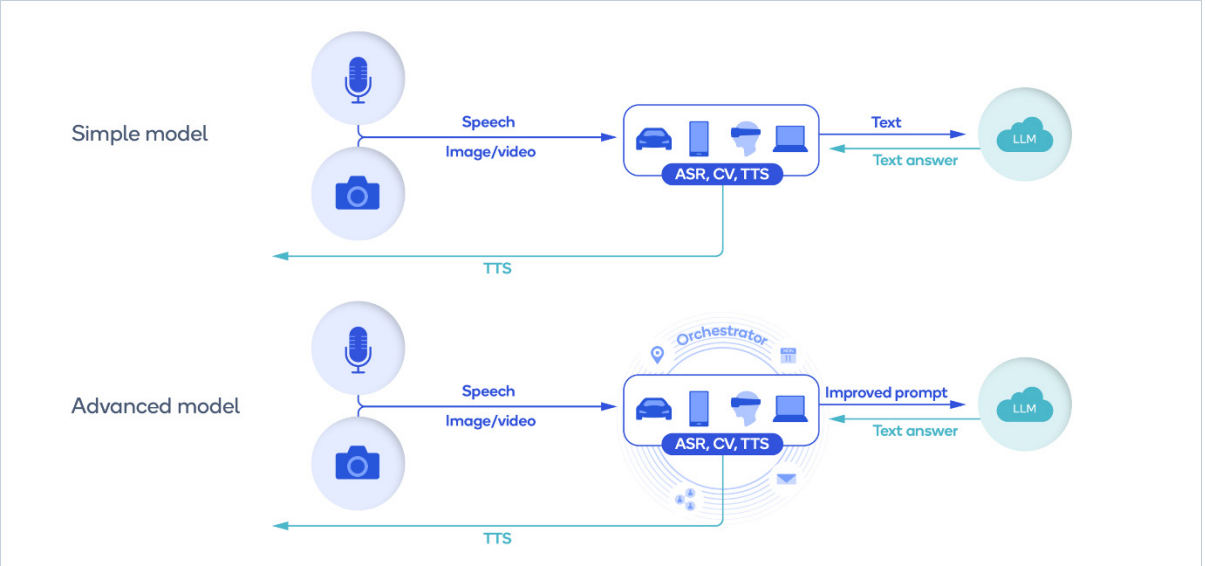


Figure 3: In device-sensing hybrid AI, the ASR, CV, and TTS happens on device. For the more advanced version, an on-device orchestrator provides an improved and more personalized prompt to the cloud.

3.3.3 Joint-processing hybrid AI

On-device and cloud AI computing can also work jointly to process AI workloads. Multi-token generation for an LLM is an example. LLMs are memory-bound, meaning that the compute hardware is more often idle while waiting for the memory data from DRAM. LLMs produce a single token per inference that is almost equivalent to a word, which means that a model like GPT-3 must read all 175 billion parameters to generate a single word. The full model then runs again to produce the next token and so on until the entire result is complete. Since memory reads are the bottleneck, it is more efficient to run several LLMs in parallel to generate multiple tokens while reading all the parameters once from DRAM. Reading all the parameters per token consumes energy and produces heat, so sharing the parameters to speculatively run LLMs in parallel on an otherwise idle compute is a net gain in performance and energy consumed.

For four-token generation, a draft LLM, which is 7-to-10 times smaller and hence less accurate than the original target LLM, runs sequentially on device four times in a row to generate the next four tokens. The device sends the four tokens to the cloud, which checks the accuracy by running the target model effectively four times with only one read of the full model parameters. The tokens are computed in parallel, with each target model having an input of zero, one, two, three, or four of the predicted tokens. These tokens are considered "drafts" until they are confirmed or corrected by the cloud. This speculative decoding process continues until the full answer is completed. Our early experiments and other published results[10] show that with four-

[10] Leviathan, Yaniv, Matan Kalman, and Yossi Matias. "Fast Inference from Transformers via Speculative Decoding." *arXiv preprint arXiv:2211.17192* (2022)

token speculative decoding, on average 2 to 3 are correct and accepted, which results in a net speedup in tokens per unit time and energy savings.

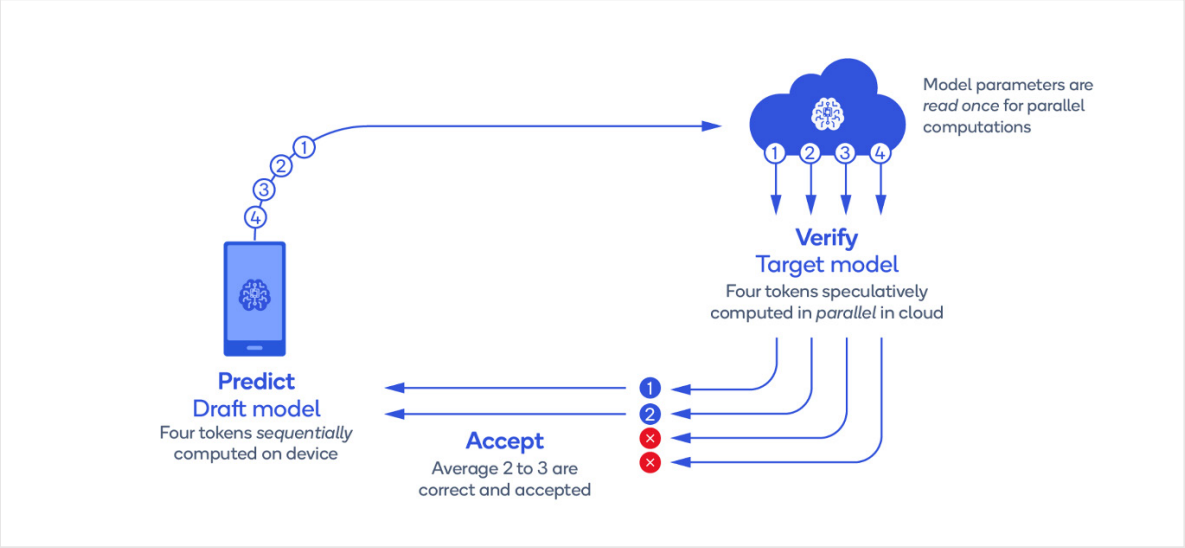


Figure 4: Four-token speculative decoding example of joint-processing hybrid AI.

4 On-device AI evolution intersects the needs of generative AI

On-device AI capabilities are key to enabling hybrid AI and allowing generative AI to reach global scale. How processing is split between the cloud and edge devices will depend on several factors, such as the capabilities of the device, privacy and security requirements, performance requirements, and even business models (see section 3.3).

AI processing has been moving toward the edge prior to the arrival of generative AI, with more and more AI inference workloads running on phones, laptops, XR headsets, vehicles, and other edge devices. For example, phones utilize on-device AI for many daily features, such as low-light photography, noise cancellation, and face unlock.

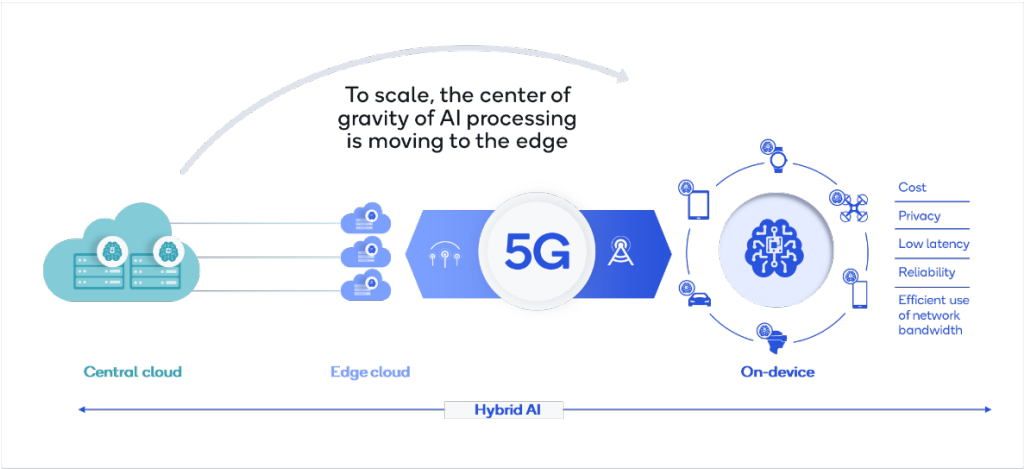
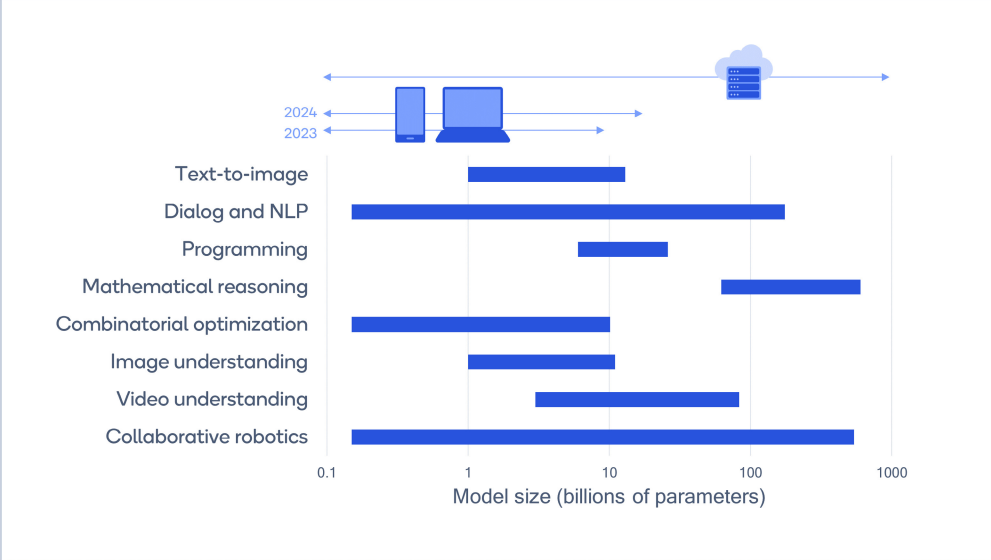


Figure 5: AI processing is gravitating toward the edge.

4.1 On-device processing can support a variety of generative AI models

With an installed base of billions of AI-capable phones, PCs, and other devices in users' hands today,[11] the potential to tap into a large amount of on-device AI processing for generative AI is already significant – and poised to grow steadily in the coming years.

A key question becomes which generative AI models can run on device with appropriate performance and accuracy. The great news is that very capable generative AI models are getting smaller while on-device processing capabilities continue to improve. Figure 2 shows a broad number of generative AI capabilities that can run on device using models that range from 1 to 10 billion parameters.[12] Models like Stable Diffusion, with over 1 billion parameters, are already running on phones with performance and accuracy levels similar to the cloud equivalent, and many other generative AI models with 10 billion parameters or more are slated to run on device in the near future.



15

Figure 6: A significant number of generative AI models can run on the device, offloading the cloud.

5 Impactful generative AI use cases across device categories

The rise of generative AI with foundation models is driving a new wave of use cases around content generation, search, and productivity across device categories, including smartphone, laptop and PC, automotive, XR, and IoT. The hybrid AI architecture will enable generative AI to provide new and enhanced user experiences across these segments.

[11] https://www.qualcomm.com/products/mobile/snapdragon/smartphones/mobile-ai
[12] Assuming INT4 parameters

5.1 Smartphone: search and digital assistant

With over 10 billion searches per day and mobile accounting for over 60% of searches,[13] the adoption of generative AI will drive a substantive increment in computing capacity required – particularly from queries made on smartphones. Users are already migrating to generative AI-based search since it provides better answers for many queries.

The popularity of chat as a search interface is also poised to increase the number of overall queries. As chat improves and becomes more capable, the smartphone can become a true digital assistant. Users can communicate naturally to get accurate and relevant answers, thanks to the combination of very accurate on-device personas and the LLMs understanding text, voice, images, video, and any other input modality. Models that do natural language processing, image understanding, video understanding, text-to-text generation, and much more will be in high demand.

5.2 Laptop and PC: productivity

With its ability to quickly generate high-quality content from simple prompts, generative AI is transforming productivity. Microsoft Office 365 on laptops and PCs is a great example. With more than 400 million Microsoft Office 365 seats and subscribers worldwide, incorporating generative AI into everyday workflows will have a significant impact.[14] Tasks that took hours or days can now take minutes. Microsoft 365 Copilot combines the power of LLMs with user data in the Microsoft Graph and the Microsoft 365 apps to turn prompts into a powerful productivity tool.[15]

Office workers can run an LLM in the background to read or compose emails in Outlook, write a document in Word, create a presentation in PowerPoint, analyze data in Excel, or collaborate in Teams meetings. Generative AI models such as natural language processing, text-to-text generation, image generation, video generation, and programming – to name a few – require tremendous amounts of processing for these heavily used productivity tasks. Most of the processing could happen on the PC in a device-centric hybrid AI architecture.

5.3 Automotive: digital assistant and autonomous driving

Informed by data in and around the vehicle, today's AI-driven cockpits offer highly personalized experiences. Similar to smartphones and PCs, in-vehicle digital assistants will keep drivers and passengers seamlessly connected with a hands-free natural user interface, while also creating new monetization opportunities for the ecosystem.

A digital assistant can access a user's personal data, such as apps, services, and payment information, as well as sensor data from the vehicle, including cameras, radar, lidar, cellular vehicle-to-everything (C-V2X), and more. Enterprise APIs also allow third-party service providers to integrate their offerings, extending their customer relationships into the vehicle. For example, the navigation experience will vastly improve with proactive assistance, such as traffic and weather updates that impact the driver's usual travel route and recommendations

[13] https://www.statista.com/statistics/297137/mobile-share-of-us-organic-search-engine-visits/
[14] Microsoft earnings reports
[15] https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/

for recharging the vehicle or purchasing a parking permit. Plus, it can order the user's favorite meal with the stored credit card with a simple ask. The cockpit media experience will also be transformed as the vehicle recognizes each occupant and customizes their experience and content, such as music or podcasts. As in-vehicle augmented reality (AR) becomes more prevalent, the digital assistant can customize the display according to the driver or passenger's preferences.

Vehicle maintenance and servicing will also become much more proactive and seamless. Analyzing data such as sensor input, maintenance history, and driving behavior, a digital assistant can predict when maintenance will be necessary. Using generative AI, the assistant can provide information on how to repair the vehicle or advise the user on finding the right service provider, improving vehicle reliability while reducing time and cost.

Advanced driver assistance systems and autonomous driving (ADAS/AD) solutions are often confused by unusual or unfamiliar objects that a perception stack has never encountered before. Typically caused by poor lighting or challenging weather conditions, this leads to unpredictable and sometimes even dangerous outcomes from the drive policy stack. To prevent similar situations in the future, the corner case data must be captured and labeled appropriately, and the model retrained. This loop can be quite laborious and time-consuming. Generative AI can create simulated corner-case scenarios predicting the trajectory and behavior of various agents on the road – such as vehicles, pedestrians, cyclists, and motorcyclists. The planner can utilize these scenarios to decide the drive policy of a vehicle.

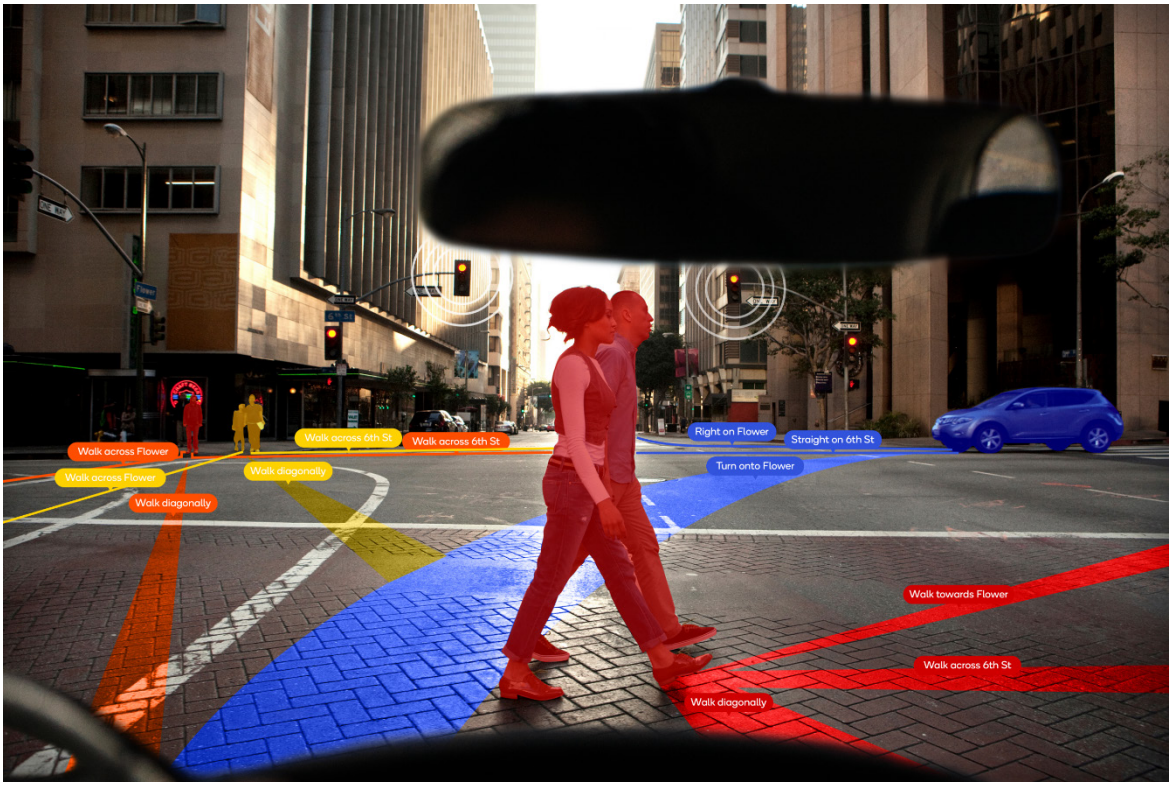


Figure 7: Generative AI can be used for ADAS/AD to help improve drive policy by predicting the trajectory and behavior of various agents.

A drive policy stack, along with a perception stack, always runs locally within the AI computing capabilities of a vehicle. Strict latency requirements preclude the cloud from playing any role in decision-making for these AI workloads. As ADAS/AD solutions embrace generative AI models with appropriate post-processing, it is inevitable that significant energy-efficient AI computing power in vehicles is required.

5.4 XR: 3D content creation and immersive experiences

Generative AI holds tremendous promise for XR. It has the potential to democratize 3D content creation and bring virtual avatars to life. The next generation of AI rendering tools will enable content creators to use a variety of prompts – such as text, speech, images, or video – to generate 3D objects, scenes, and (eventually) entire virtual worlds. In addition, content creators will get to leverage text-to-text LLMs to generate human-like conversations for avatars that are fully voiced and emotive. Taken together, these advancements will transform the way we create and experience immersive content on XR devices.

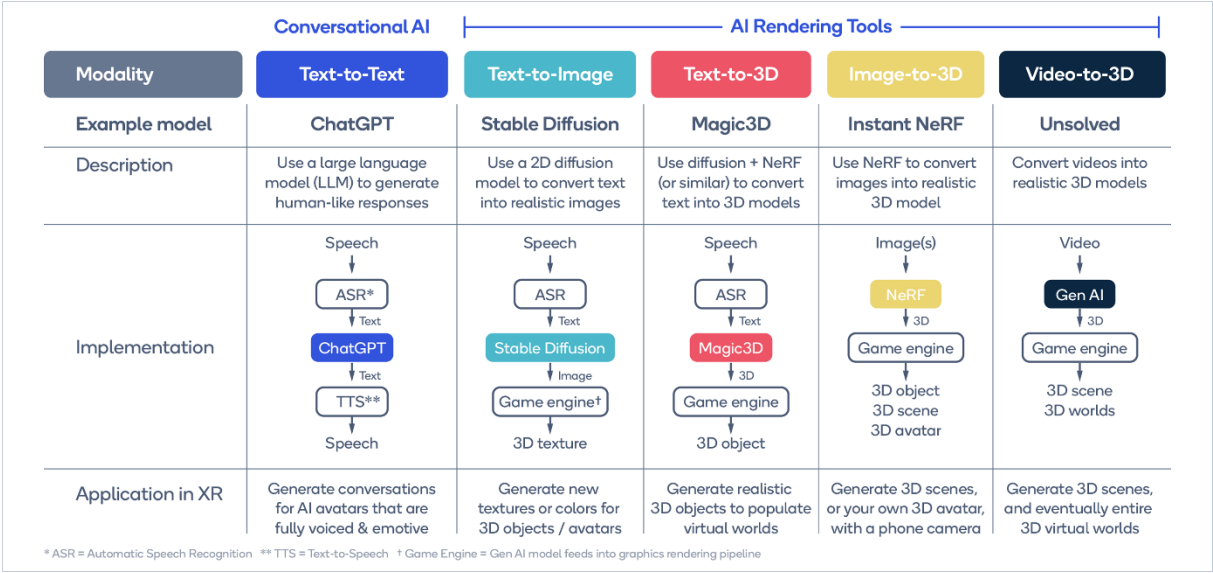


Figure 8: Generative AI models will enable conversational AI and new rendering tools for XR.

While the promise of generative AI for XR is undoubtably exciting, it is difficult to predict when there will be widespread adoption of these technologies. However, given the rapid pace of innovation in recent months, it is safe to say that we can expect to see significant progress in the coming years.

For immersive worlds, text-to-image models, such as Stable Diffusion, will soon enable content creators to generate realistic textures for 3D objects. We expect these capabilities to become available on smartphones, and by extension XR devices, within a year. Deployment in XR will require "distributed processing," where the headset runs the perception and rendering stack, while the paired smartphone or the cloud run the generative AI model. In a couple of years, the first text-to-3D and image-to-3D models will likely make it to the edge to generate high-quality point clouds of 3D objects. Within a few years later, these models will improve further, to the point where they can generate high-quality 3D textured objects from scratch. In a decade or so, further model improvements will enable entire 3D rooms and scenes to generate, at high

fidelity, from text or images. In the future, text-to-3D and video-to-3D models might finally allow us to step into 3D virtual worlds that were generated from scratch, for example automatically constructing a 3D virtual environment that is only limited by the user's imagination.

Figure 9: Generative AI will help create immersive 3D virtual worlds based on simple prompts, such as: "surreal world, floating jellyfish all around, beautiful waterfalls, mystical lake, towering mountains."

Virtual avatars will follow a similar progression. Text-to-text models – such as 13 billion parameter LLaMA – will make it to edge devices, to generate conversations for avatars that feel natural and intuitive. In addition, text-to-image models will generate new textures and outfits for these avatars. In the following years, image-to-3D and encoder/decoder models will generate head and full-body avatars of humans for telecommunication. Eventually, we will use voice prompts, images, or video to generate human-like virtual avatars that are photorealistic, fully animated, intelligent, and mass producible.

5.5 IoT: Operational efficiency and customer support

AI is already used today in a wide variety of IoT verticals, including retail, security, energy and utilities, supply chain, and asset management. AI improves decision-making based on data collected and analyzed in near real time, optimizes operational efficiency, and enables innovation to create competitive differentiation. IoT segments stand to further benefit from AI utilization through generative AI.

For example, generative AI can improve the customer and employee experience in retail. A grocery shopping agent at an onsite kiosk or smart cart can make a menu with recipes using the weekly sales specials, budget constraints, and family preferences. Store managers can anticipate off-cycle sales opportunities based on upcoming events and prepare accordingly. If a sports team was coming into town, a store manager can ask generative AI what branded items

fans like and increase inventory accordingly. Another use is to create a new store planogram based on best practices and positive results from other stores in similar communities. Generative AI could help store managers reorder products in shelves to expand space for the most profitable brands using a simple prompt, or it could help minimize the prominence of out-of-stock items on the shelf using data from nearby chain stores.

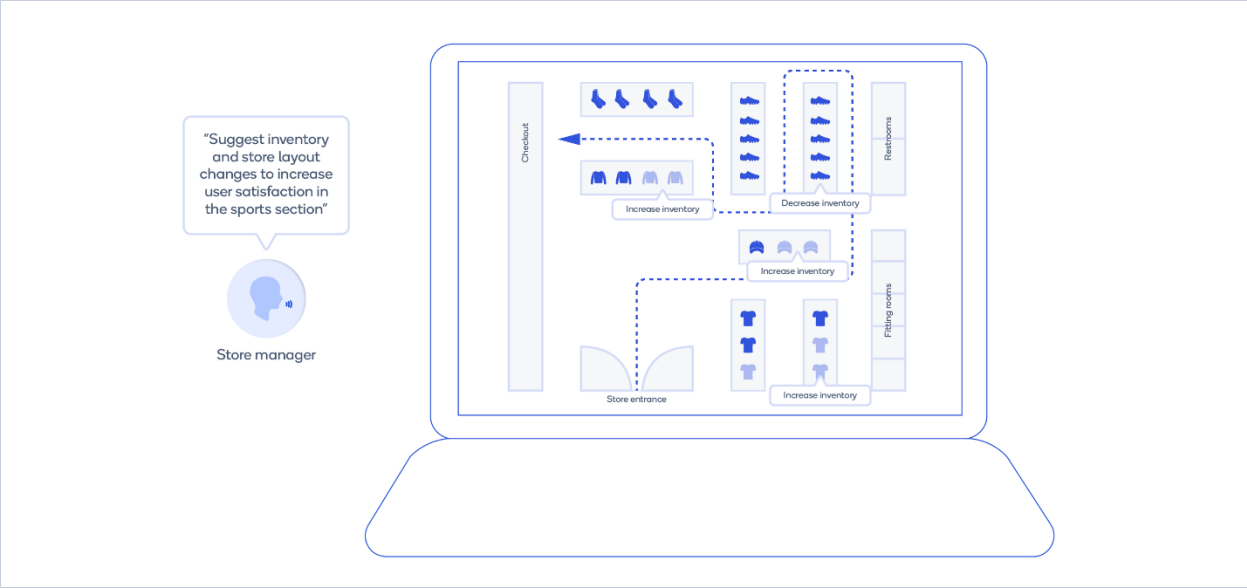


Figure 10: Generative AI will help improve customer and employee experience in retail, such as providing recommendations for inventory and store layout.

The energy and utilities sector also stands to benefit from generative AI. Operation teams can create corner-case load scenarios and predict demand for electricity along with potential grid failures under unusual circumstances, such as a hot summer with strong winds and localized fires in rural areas, to better manage their resources and avoid outages. Generative AI can also be used to provide better customer service, such as by answering questions about outages or billing.

6 Conclusion

Hybrid AI is inevitable. Use cases for generative AI will continue to evolve and grow into mainstream experiences, increasing demand on the cloud and its infrastructure. With the advanced capabilities of on-device AI, a hybrid AI architecture can scale to meet enterprise and consumer needs, providing benefits with regards to cost, energy, performance, privacy, security, and personalization. The cloud and devices will work together to deliver next-generation user experiences through powerful, efficient, and highly optimized AI capabilities.

Qualcomm

Follow us on:
For more information, visit us at:
qualcomm.com