Qualcomm

AI disruption is driving innovation in on-device inference

How the proliferation and evolution of generative models will transform the AI landscape and unlock value.

February 2025

Contents

Executive summary 3
Quality AI models are now abundant and affordable 4
Innovations boost model quality and reduce development time and cost 4
Small models achieve big capabilities at the edge 5
The era of AI inference innovation is here 7
Qualcomm is set to be a leader in the AI inference era 8
Expanding across all key edge segments 9
Mobile 9
PCs 10
Automotive 10
Industrial IoT 11
Networking 11
Conclusion 11

Executive summary

The introduction of DeepSeek R1, a cutting-edge reasoning AI model, has caused ripples throughout the tech industry. That's because its performance is on par with or better than state-of-the-art alternatives, disrupting the conventional wisdom around AI development.

This pivotal moment is part of a broader trend that underscores the innovation in creating high-quality small language and multimodal reasoning models, and how they're preparing AI for commercial applications and on-device inference. The fact that these new models can run on devices accelerates scale and creates demand for powerful chips at the edge.

Driving this shift are four major trends that are leading to a dramatic improvement in the quality, performance, and efficiency of AI models that can now run on device:

- **Today's state-of-the-art smaller AI models have superior performance.** New techniques like model distillation and novel AI network architectures simplify the development process without sacrificing quality, allowing new models to outperform larger ones from a year ago, which could only operate on the cloud.
- **Model sizes are decreasing rapidly**. State-of-the-art quantization and pruning techniques allow developers to reduce the size of models with no material impact in accuracy.
- **Developers have more to work with.** The rapid proliferation of high-quality AI models means features like text summarization, coding assistants and live translation are common in devices like smartphones, making AI ready for commercial applications at scale across the edge.
- **AI is becoming the new user interface.** Personalized multimodal AI agents will simplify interactions and proficiently complete tasks across various applications.

Qualcomm Technologies is strategically positioned to lead and capitalize on the transition from AI training to large-scale inference, as well as the expansion of AI computational processing from the cloud to the edge. The company has an extensive track record in developing custom central processing units (CPUs), neural processing units (NPUs), graphics processing units (GPUs), and low-power subsystems. The company's collaboration with model makers, along with tools, frameworks, and SDKs for deploying models across various edge device segments, enables developers to accelerate the adoption of AI agents and applications at the edge.

The recent disruption and reassessment of how AI models are trained validates the imminent AI landscape shift towards large-scale inference. It will create a new cycle of innovation and upgrade of inference computing at the edge. While training will continue in the cloud, inference will benefit from the scale of devices running on Qualcomm® technology and create demand for more AI-enabled processors at the edge.

Quality AI models are now abundant and affordable

Innovations boost model quality and reduce development time and cost

AI has reached the point where the drop in the cost of training AI models, combined with open-source collaboration, is making the development of high-quality models accessible to more people and organizations.

This shift is driven by various technical advancements. Usage of longer context length, along with simplification of some of the training steps, saves computational costs. Newer network architectures ranging from mixture-of-experts (MoE) to state-space models (SSM) are pushing the boundary of what can be accomplished with reduced computational overhead and power consumption.

Newer AI models also integrate advanced methods such as chain-of-thought reasoning and self-verification, enabling them to perform well across various challenging domains like mathematics, coding, and scientific reasoning.

Distillation is a key technique in the development of capable small models. It allows large models to "teach" smaller models, transferring knowledge while maintaining accuracy. The use of distillation has led to a surge in smaller foundation models—many of them fine-tuned for specialized tasks.

The power of distillation is exemplified in figure 1. This presents average LiveBench results comparing the Llama 3.3 70B model with its distilled DeepSeek R1 counterpart. The chart shows how distillation significantly enhances performance in reasoning, coding, and mathematics tasks for the same number of parameters.

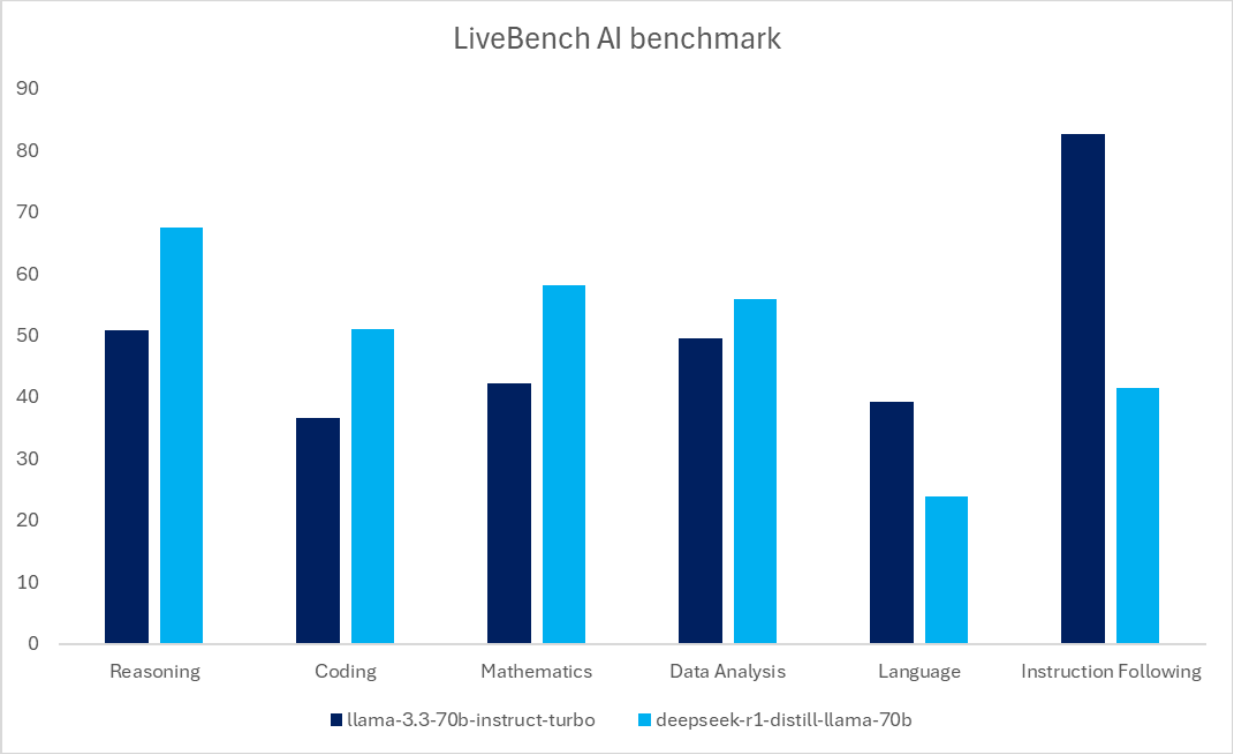

Figure 1: LiveBench AI average benchmark results comparing Meta Llama 70B model with its distilled counterpart by DeepSeek. Source: LiveBench.ai, Feb. 2025.

Small models achieve big capabilities at the edge

Smaller models are approaching the quality of large frontier models due to distillation and other techniques described above. Figure 2 shows benchmarks for the DeepSeek R1 distilled models compared to leading-edge alternatives. DeepSeek-distilled versions based on Qwen and Llama models show areas of significant superiority, particularly in the GPQA benchmark – achieving superior or similar scores compared to state-of-the-art models such as GPT-4o, Claude 3.5 Sonnet, and GPT-o1 mini. GPQA is a critical metric because it involves deep, multi-step reasoning to solve complex queries, which many models find challenging.

Model	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	44.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633

Source: https://github.com/deepseek-ai/DeepSeek-R1

Figure 2: Mathematic and coding benchmarks. Source: DeepSeek, Jan. 2025.

Many popular model families including DeepSeek R1, Meta Llama, IBM Granite, Mistral Ministral feature small variants which overdeliver in terms of performance and benchmarks for specific tasks, regardless of their size. The reduction of large, foundational models into smaller, efficient versions enables faster inference, smaller memory footprint and lowers power consumption – all while maintaining a high bar on performance, allowing deployment of such models within devices like smartphones, PCs, and automobiles.

Further optimizations, like quantization, compression and pruning help reduce model sizes. Quantization lowers power consumption and speeds up operations by reducing precision without significantly sacrificing accuracy, while pruning eliminates unnecessary parameters.

These technical developments have led to a proliferation of high-quality generative AI models. According to data compiled by Epoch AI (Figure 3), more than 75% of large-scale AI models published in 2024 feature less than 100 billion parameters.

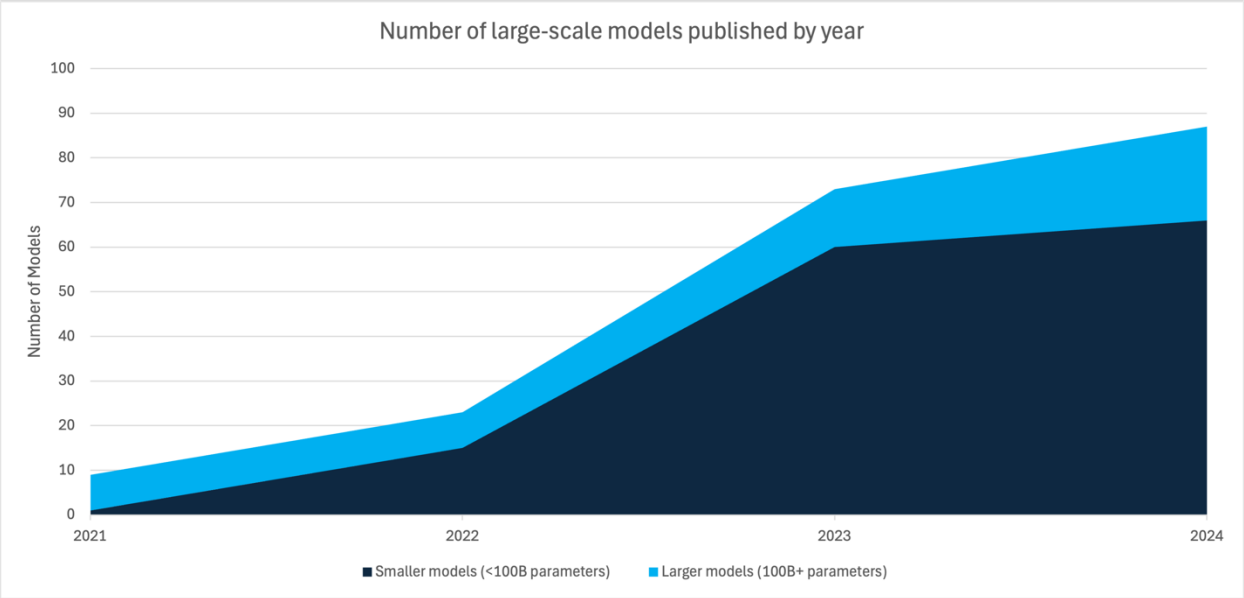

Figure 3: Number of large-scale AI models published by year, categorized by number of parameters. Source: Epoch AI, Jan. 2025.

The era of AI inference innovation is here

The abundance of high-quality, smaller models is bringing renewed attention to inference workloads – which is where applications and services make use of the models to provide value to businesses and consumers.

Qualcomm Technologies has worked on the optimization of numerous AI models to support the commercialization of the new generation of AI-oriented Copilot+ PCs. Similarly, the company has collaborated with OEMs such as Samsung and Xiaomi in the launch of flagship smartphones equipped with many AI-enabled features.

The proliferation of AI inferencing capabilities across devices has enabled the creation of generative AI applications and assistants. Document summarization, AI-image generation and editing, and real-time language translation are now common features. Camera apps leverage AI for computational photography, object recognition and real-time scene optimization.

Next up is the development of multimodal applications which combine multiple types of data—text, vision, audio and sensor input—to deliver richer, more context-aware and personalized experiences. The Qualcomm AI Engine combines the capabilities of custom-built NPUs, CPUs and GPUs to optimize such tasks on-device, enabling AI assistants to switch between communication modes and generate multimodal outputs.

Agentic AI is positioned at the heart of the next generation of user interfaces. AI systems

are capable of decision-making and task management by predicting user needs and proactively executing complex workflows within devices and applications. Qualcomm Technologies' emphasis on efficient, real-time AI processing allows these agents to function continuously and securely within the devices, while relying upon a personal knowledge graph that accurately defines the user's preferences and needs, without any cloud dependency. Over time, these advancements are laying the groundwork for AI to become the primary UI, with natural language and image, video and gesture-based interactions simplifying how people engage with technology.

Looking ahead, Qualcomm Technologies is also positioned for the era of embodied AI, in which AI capabilities are integrated into robotics. By leveraging its expertise in inference optimization, Qualcomm Technologies aims to power real-time decision-making for robots, drones and other autonomous devices, enabling precise interactions in dynamic, real-world environments.

While numerous AI models are trained in the cloud, distilled smaller models are available for operation and run on devices often within weeks or days. For example, within less than a week, DeepSeek R1-distilled models were running on PCs and smartphones powered by Snapdragon® platforms.

Deploying inference within devices addresses immediacy through reduced latency, enhances privacy, relies on local data to provide additional context and enables continuous functionality of AI features and applications. It also reduces costs for users and/or developers by avoiding fees associated with cloud inference services. All of this creates incentives for software and service providers to deploy AI inference at the edge.

Qualcomm is set to be a leader in the AI inference era

As a leader in on-device AI, Qualcomm Technologies is strategically positioned to advance the AI inference era with its industry-leading hardware and software solutions for edge devices. These solutions encompass billions of smartphones, automobiles, XR headsets and glasses, PCs, industrial IoT devices, and more.

Qualcomm Technologies has a long history of developing custom CPUs, NPUs, GPUs and low-power subsystems, which, when combined with expertise in packaging and thermal design, form the foundation of its industry-leading system-on-chip (SoC) products. These SoCs deliver high-performance, energy-efficient AI inference directly on-device. By tightly integrating these cores, Qualcomm Technologies' platforms can handle complex AI tasks while maintaining battery life and overall power efficiency—critical for edge use cases.

To unlock the full potential of AI on its platforms, Qualcomm Technologies has built a robust AI software stack designed to empower software developers. The Qualcomm AI

Stack includes libraries, SDKs, and optimization tools that streamline model deployment and enhance performance. Developers can leverage these resources to efficiently adapt models for Qualcomm platforms, reducing time-to-market for AI-powered applications. Qualcomm Technologies' developer-focused approach accelerates innovation by simplifying the integration of cutting-edge AI features into consumer and enterprise products.

Lastly, the company's collaboration with AI model makers across the globe and its provision of services like the Qualcomm AI Hub are central to its strategy for scaling AI across industries. On the Qualcomm AI Hub, in three simple steps, a developer can 1) pick a model or bring their own model or create a model based on their data; 2) pick any framework and runtime, write and test their AI apps on a cloud-based physical device farm; and 3) use tools to deploy their apps commercially. The Qualcomm AI Hub supports major large language and multimodal model (LLM, LMM) families, allowing developers to deploy, optimize, and manage inference on devices powered by Qualcomm platforms. With features like pre-optimized model libraries and support for custom model optimization and integration, Qualcomm Technologies enables rapid development cycles while enhancing compatibility with diverse AI ecosystems. This collaborative approach strengthens Qualcomm Technologies' position as a leader in enabling scalable, real-time AI applications.

Expanding across all key edge segments

Qualcomm Technologies uses on-device AI to support many industries, unlocking business value and supporting new user experiences, all enabled by enhanced performance, efficiency, responsiveness and privacy by processing AI locally on devices.

Mobile

Snapdragon mobile platforms, such as the latest Snapdragon 8 elite, are advancing the capabilities of on-device AI by enabling several cutting-edge multimodal generative models and agentic AI to operate natively on smartphones. AI has enhanced smartphone features across various categories such as communication improvement, generative image editing tools, personalization, and accessibility. On-device generative AI is being utilized to develop more intuitive, user-centric features and to automate tasks in mobile devices.

This trend towards AI-driven functionalities is evident in the latest flagship smartphone releases from major manufacturers utilizing Snapdragon platforms, including Samsung, ASUS, Xiaomi, Oppo, Vivo, and Honor.

PCs

Snapdragon X Series platforms were instrumental in defining the new category of AI PCs, with best-in-class custom NPU cores that were built from ground-up for high performance, energy efficient generative AI inference. This NPU is turbo-charging Windows apps, adding new features, boosting performance, and enhancing privacy and battery life. Developers can run generative AI inference on-device, offering cutting-edge Copilot+ PC features which debuted on the Snapdragon X Series.

Popular third-party apps like Zoom, Affinity, Djay Pro, CapCut, Moises Live, and Blackmagic Design's DaVinci Resolve take advantage of the NPU to offer specific AI-powered capabilities on Snapdragon X Series platforms.

Automotive

Snapdragon® Digital Chassis™ solution uses on-device AI in its context-aware intelligent cockpit system designed to enhance vehicle safety and driver experience. This system leverages advanced cameras, biometric and environmental sensors, and state-of-the-art multimodal AI networks to provide real-time feedback and functionality tailored to the driver's state and environmental conditions.

For automated driving and assistance systems, Qualcomm Technologies has developed an end-to-end architecture which uses large training datasets, fast re-training using real-world and AI-augmented data, over-the-air updates, and a state-of-the-art stack including multimodal AI models and causal reasoning in the vehicle to handle modern automated driving and assistance complexities.

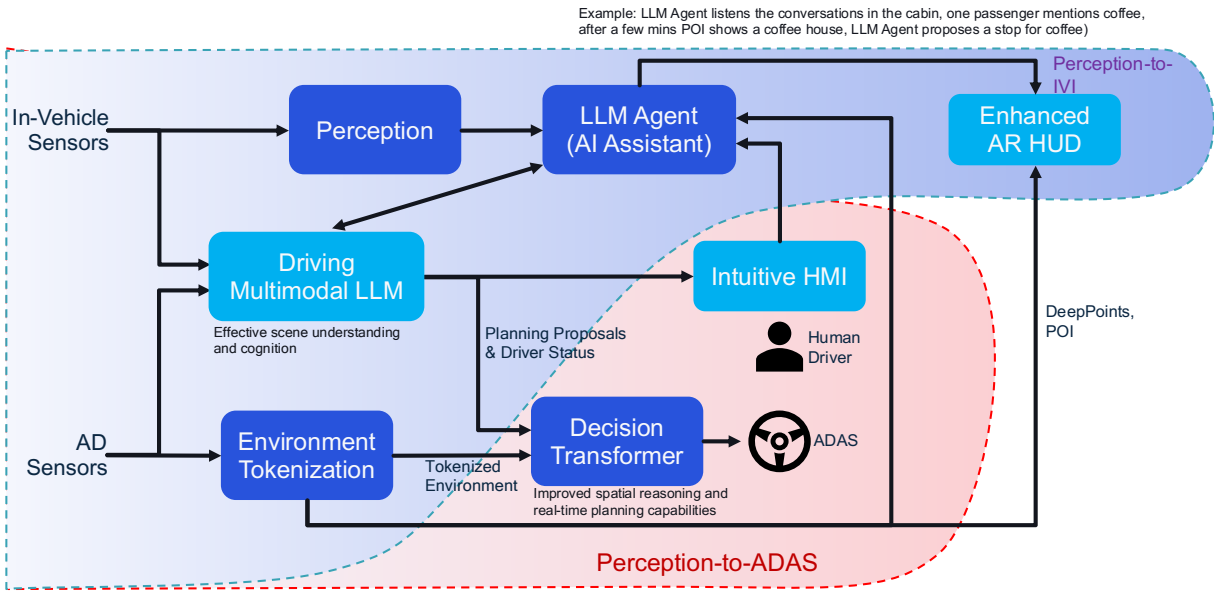

Figure 4: Simplified in-vehicle AI system architecture to support intelligent cockpit and autonomous and advanced driving assistance. Source: Qualcomm Technologies, Jan. 2025,

Industrial IoT

For industrial IoT and enterprise applications, Qualcomm Technologies recently introduced its the Qualcomm® AI On-Prem Appliance Solution, an on-premises desktop or wall-mounted hardware solution, and Qualcomm® AI Inference Suite, a set of software and services for AI inferencing spanning from near-edge to cloud.

This edge AI approach allows sensitive customer data, fine-tuned models, and inference loads to remain on premises, enhancing privacy, control, energy efficiency, and low latency. That's critical for AI-enabled business applications such as intelligent multi-lingual search, custom AI assistants and agents, code generation, and computer vision for security, safety and site monitoring.

Networking

Qualcomm Technologies has introduced an AI-enabled Wi-Fi networking platform – the Qualcomm® Networking Pro A7 Elite. The solution integrates Wi-Fi 7 and edge AI to allow access points and routers to run generative AI inference on behalf of connected devices in the network. It supports innovative applications in areas like security, energy management, virtual assistants, and health monitoring by processing data on the gateway for enhanced privacy and real-time responses.

This networking platform is expected to transform Wi-Fi routers, mesh systems, broadband gateways, and access points into private, local AI-based mini-servers within homes and enterprises.

Conclusion

AI is undergoing a transformative shift driven by falling training costs, rapid inference deployment, and innovations tailored to edge environments. The tech industry focus is no longer dominated by the race to build larger models, but by efforts to efficiently deploy them in real-world applications at the edge.

The distillation of large foundation models has unleashed a surge of smarter, smaller, and more efficient models, empowering industries to integrate AI faster and at scale – increasingly within devices themselves.

Qualcomm Technologies is uniquely positioned to lead and benefit from this change through its expertise in power-efficient chip design, advanced AI software stack, and comprehensive developer support for edge applications.

Qualcomm Technologies' ability to integrate NPUs, GPUs, and CPUs into devices enables high-performance, energy-efficient AI inference across smartphones, PCs, automotive, and industrial IoT sectors. Qualcomm Technologies provides industries with high performance, affordable, responsive, and privacy-oriented transformative AI experiences.

The company's ecosystem approach—encompassing its Qualcomm AI Stack, Qualcomm AI Hub, and strategic developer collaborations—accelerates the deployment of adaptive AI technologies. These solutions help meet the demands of industries prioritizing real-time performance, privacy, and efficiency.

As AI innovation explodes at the edge, Qualcomm Technologies' investments in scalable hardware and software will further solidify its leadership. The company is enabling a new era where AI applications are more accessible, efficient, and integrated into everyday life, driving transformation across multiple sectors globally.