NEWS RELEASE

Ambiq Launches heliaAOT™ 1.0 Delivering Up to 76% Lower Edge AI Inference Latency

2026-11-09

Up to 72% lower energy per inference and broader model support help customers bring more capable AI to Apollo-based devices within existing power and memory constraints

AUSTIN, Texas--(BUSINESS WIRE)-- Ambiq® Micro, Inc. (“Ambiq”), a technology leader in ultra-low-power semiconductor solutions for edge AI, today introduced heliaAOT 1.0, its ahead-of-time AI compiler for the HELIA™ AI ecosystem. This development further highlights Ambiq’s focus on building an integrated silicon and software platform that enables Edge AI compute on low-power, memory-constrained devices.

heliaAOT converts trained AI models into optimized C code, reducing processing time, energy, and memory required for on-device inference. In benchmark tests on Apollo510, heliaAOT delivered up to 76% lower inference latency and up to 72% lower energy per inference than LiteRT Micro[1]. heliaAOT has already been deployed in mass-produced products, and Ambiq is in discussions with some of its larger customers about deploying it in their current products.

“Customers want to bring more intelligence to their devices without exceeding their power and memory budgets,” said Dr. Adam Page, Head of Artificial Intelligence at Ambiq. “heliaAOT 1.0 helps them get more from Apollo silicon, deploy a broader range of models, and reuse their development work across products.”

More AI Capability Across Apollo Families

heliaAOT 1.0 helps developers run more models, use less memory, and reuse their work across Apollo products:

- Broader model support: Run supported FP16 and FP32 models without integer quantization, with approximately 4× faster FP32 inference than LiteRT Micro FP32 in benchmark testing[2]. Support for 65 operators plus custom operators accommodates more model architectures, enabling more AI use cases.
- Smaller footprint: Compared with LiteRT Micro, median firmware images are 29% to 56% smaller, freeing up storage and RAM for more applications[3].
- Always-on intelligence: Streaming LSTM, GRU, and SVDF support enables continuous voice, biosignal, and sensor processing for greater device responsiveness.
- Easier deployment: Retarget supported models across 15 Apollo SoCs with a one-line compiler configuration change, saving development time.

In Apollo510 benchmarks against LiteRT Micro, heliaAOT reduced inference latency, energy consumption, and firmware image size across all tested workload groups.

AI workload	Lower Latency	Lower energy per inference	Smaller firmware image
Health monitoring: ECG, PPG, and sleep	76%	72%	56%
Vision	61%	51%	29%
Voice and audio	52%	45%	47%
MLPerf Tiny reference models	36%	27%	37%

Ambiq tests every release on Apollo silicon. With heliaPROFILER™, developers can measure performance on their own models and boards to assess the benefits for their applications.

Advancing the Apollo AI Ecosystem

Ambiq delivered 17 heliaAOT updates over the past 12 months, expanding operator support from 35 to 65 and Apollo target support from 8 to 15. Recognized as the Best Tools category at Embedded World 2026, heliaAOT strengthens Ambiq's combined silicon and software offering and helps customers carry their AI development work across products.

heliaAOT 1.0 is available now. Explore the benchmarks and request an evaluation at **https://ambiq.com/ai/heliaaot/**.

About Ambiq

Ambiq's mission is to enable intelligence (artificial intelligence (AI) and beyond) everywhere by delivering the lowest power semiconductor solutions. Built on its patented Subthreshold Power Optimized Technology (SPOT®) platform and the HELIA™ AI ecosystem, Ambiq empowers manufacturers to bring more capable AI to the edge, where power,

memory, and energy efficiency are most critical. Ambiq enables more intelligent, always-on edge devices across healthcare, wearables, industrial automation, smart environments, and other emerging AI applications. With more than 300 million devices shipped worldwide, Ambiq continues to shape the future of always-on Edge AI. Headquartered in Austin, Texas, Ambiq serves customers globally. For more information, visit **www.ambiq.com**.

Ambiq, Ambiq Micro and the Ambiq logo are registered trademarks of Ambiq Micro, Inc. All other company or product names noted herein may be trademarks of their respective holders.

Availability of Information on Our Website

We routinely use our investor relations website (ir.ambiq.com) to post presentations to investors and other important information, including information that may be material. Accordingly, we encourage investors and others interested in Ambiq to review the information we make public on our investor relations website.

Foot Notes

---

[1] Measured on Apollo510 at 96 MHz against LiteRT Micro under matched per-model conditions. Energy is gross MCU-rail energy per inference. Results are model-level, not battery-life, measurements. MLPerf Tiny refers to the reference models and is not an official MLPerf submission.
[2] Matched-FP32 comparison across six model families on Apollo510.
[3] Median reductions versus LiteRT Micro. Firmware image size is flashed storage, and model buffers exclude stack and heap. Neither is total peak RAM.

Company Contact:

Charlene Wan
VP of Corporate Marketing
P: +1-512-879-2850
E: **cwan@ambiq.com**

Investor Relations Contact:

Teneo
Christina Coronios
E: **christina.coronios@teneo.com**

Source: Ambiq Micro, Inc.