Nov 13, 2026

Generative EDGE AI WG

Invited roberto.morabito@eurecom.fr davis.sawyer@nxp.com Evgeni Gousev danilo.pau@st.com msingh@qti.qualcomm.com ashutosh2.kumar@intel.com raashid.ansari@silabs.com mkstaufferpa@gmail.com simplici@qti.qualcomm.com Rosina Haberl danilopietro.pau@gmail.com m.rueb@foresthub.ai michael.stauffer@alumni.stanford.edu tinoosh@jhu.edu niall.lyons@infineon.com kiruba.subramani@silabs.com ziereis@roofline.ai omar.shrit.zj@renesas.com marcus.rueb@hahn-schickard.de bernd.hollensteiner@renesas.com ashutosh.pandey@infineon.com tara.khani@edgeaiinnovations.com

Attachments Generative EDGE AI WG

Meeting records Transcript

Summary

Workshop planning focused on small language model edge benchmarking and December 14th scheduling.

Workshop Proposals and Scheduling
Workshop proposals require approval for specific kadas costs. December 14th was selected for the event to accommodate European time zones.

Format and Speaker Lineup
Events will mirror previous initiatives focusing on performance benchmarks on the edge. External speakers and hybrid structures will broaden the series.

Benchmarks and Robotics Use Cases
Standardization discussions addressed capability measurements alongside practical retail robotics use cases. Future roadmaps lead toward Embedded World in March.

Decisions

Aligned

- **Workshop event date set for December 14th** The participants agreed to schedule the upcoming workshop event for December 14th.
- **December workshop topic centered on small language model benchmarking** The group aligned on focusing the December workshop topic on measuring capabilities and performance benchmarks of small language models on the edge.

Next steps

- ☐ [Tara Khani] Write the one-pager: outlining the workshop format, topics, and schedule for the December 14th event on measuring capabilities of small language models on the edge, and share it with Pete Bernard and Danilo Pau for feedback.
- ☐ [Tara Khani] Notify the team regarding CES attendance: to inform Pete and others if plans are finalized to attend the event.
- ☐ [Pete Bernard] Invite David Caner: to speak at the December workshop and share a perspective from the ML Perf world.

Details

Workshop Proposals and Processor Options: Name not available discusses two proposed workshops, noting that approval from Ashuttosh is required for four versions of kadas that cost $1,000 to avoid confusion. The first workshop is planned by Tara Khani and Ashuttosh using Intel processors, while the second workshop involves a collaborating professor focusing on an optimized implementation on STMPU to achieve language model semantic navigation and segment navigation for robots

(00:00:01).

Workshop Scheduling and Timing: Tara Khani notes that Ashush suggested the second or third week of December, though the third week is ruled out because people tend to tune out past December 15th. Pete Bernard and Name not available agree to schedule the workshop for Monday morning, December 14th, to accommodate European time zones at 5:00 PM, with Pete Bernard noting that approximately two weeks of promotion will be needed (00:01:47).

Workshop Format and Benchmark Focus: Tara Khani explains that the event will mirror previous initiatives done with Arduino Uno Q, VentunoQ, and Snapdragon, including a blog on the Qualcomm developer hub and expanding to Intel core laptops and edge devices (00:02:44). Tara Khani proposes focusing the initial talks on performance and search engine benchmarks on the edge, highlighting a tiny footprint on the CPU using Uch compared to HNSW while achieving higher accuracy and speed (00:03:42). Pete Bernard agrees that benchmarking small language models on the edge serves as a strong pitch for viewers (00:04:46).

Expanding Speaker Lineup and Event Tracks: Pete Bernard suggests inviting external speakers from RunIQ and David Caner from ML Perf to build a broader series of talks around small language model capabilities on the edge. Name not available proposes a hybrid event structure to accommodate both hands-on sessions and frontal presentations (00:05:53). Tara Khani outlines two distinct target audiences and tracks: an agentic track for developer laptops and tiny edge devices like UNOQ, and an industrial track for AMR and retail use cases utilizing infrastructure like Qualcomm gen, CUDA, Intel Xeon, and customized layers like SEC with robotic integrators such as Seiko (00:06:56).

Event Structure, Future Roadmap, and Date Confirmation: Pete Bernard proposes running three to four talks from 8:00 AM to 12:00 PM, considering whether to use a call for papers or handpick presenters (00:09:02). Tara Khani details a long-term roadmap leading to Embedded World in March with the AGI Foundation, featuring preceding webinars and contests, while Pete Bernard suggests adding hands-on workshops and

contests at Edgi Munich at the end of that week (00:09:56). Pete Bernard formally locks in December 14th for a four-hour event focused on measuring capabilities of small language models on the edge, and Tara Khani commits to writing a one-pager immediately following the meeting to share with Pete Bernard and Name not available (00:10:56).

Standardizing Benchmarks and Proposed Fireside Chat: Pete Bernard highlights the current lack of standards for capability and performance, describing the space as "the wild west" and comparing it to similar discussions in the neuromorphic group regarding neurobench (00:11:42). Pete Bernard volunteers to ask David Caner to share a perspective from the ML Perf world, which Name not available supports as a fireside chat format to debate standardization versus performance measurement (00:12:26). Pete Bernard notes that David Caner would serve as a strong, neutral industry draw for the event (00:13:26).

Software Footprint and Robotics Use Cases: Tara Khani outlines an upcoming meeting with Ashish and their software team to test hardware and benchmarks, and compares the large GPU footprint of the Nvidia-funded Milvus vector database with a CPU benchmark using Milvus that achieved higher accuracy and lower latency with a tiny footprint, keeping NPU and GPU resources free for large language models (00:13:26). Tara Khani describes a retail robotics use case combining object detection via YOLO with a voice agent to manage inventory and staff status (00:14:21). Pete Bernard emphasizes the importance of driving the technology with practical scenarios under the theme of "small language models in action" (00:15:30).

Meeting Attendance, Communication Channels, and CES Discussion: Pete Bernard questions why only three out of twenty-two invitees attended the meeting, and Name not available attributes the low turnout to rescheduling confusion caused by travel (00:15:30). Pete Bernard introduces Joanne as the new master scheduler replacing Rosina. In response to Tara Khani, Pete Bernard encourages participants to post open-source papers, product information, and partnership requests in the WhatsApp group and utilize the career group (00:16:29). Finally, Tara Khani inquires about CES and a hack bounty invite from Ginger, prompting Pete Bernard to advise caution regarding CES due to its high cost and noise unless meetings are carefully

planned in advance (00:17:16).

You should review Gemini's notes to make sure they're accurate. Get tips and learn how Gemini takes notes

*How is the quality of **these specific notes?** Take a short survey to let us know your feedback, including how helpful the notes were for your needs.*