Generative Edge AI Working Group: Enabling Creativity and Intelligence at the Network's Edge

The Generative Edge AI Working Group is a collaborative initiative within the EDGE AI FOUNDATION dedicated to advancing the frontiers of generative artificial intelligence in real-time, resource-constrained, and decentralized environments.

As large-scale generative models continue to reshape how we interact with technology—from multimodal assistants and real-time translation to autonomous systems and industrial monitoring—bringing generative capabilities to the edge is the next bold step in AI democratization. This working group brings together academic researchers, industry practitioners, and open-source contributors to make this vision a reality.

We aim to empower edge devices with generative AI capabilities that are energy-efficient, privacy-preserving, responsive, and autonomous, unlocking intelligent behavior closer to the user, the sensor, and the moment of interaction.

Charter

Generative Edge AI is defined as a breakthrough field in edge AI and tinyML. It targets **resource-restricted generative artificial intelligence technologies and applications** including hardware, algorithms, tools, ecosystems, applications, and software solutions capable of enabling natural interaction on edge devices at extremely high energy efficiency, **typically in the peta-operations per Watt (POp/W) range or higher**.

This new field is poised to enable an unprecedented generation of **powerful yet low-power neural processor units (NPUs)**, **in-memory computing**, and **systems-on-chip (SoCs)** that leverage **heterogeneous integration** to support scalable and sustainable edge intelligence.

Definition

Generative Edge AI refers to deploying and running generative AI models directly on edge devices (e.g., smartphones, IoT devices, sensors, autonomous vehicles) rather than relying on centralized cloud infrastructure. These models generate outputs such as text, images, or actions in real time, at the point of data collection or user interaction, enabling low-latency, personalized, and private AI services.

Mission Statement

The Generative Edge AI Working Group empowers and connects academia, industry, and individuals to advance knowledge, collaboration, and innovation in Edge AI through education, community engagement, and recognition of groundbreaking achievements.

Objectives

To fulfill its mission, the Generative Edge AI Working Group has defined a set of objectives. These are designed to promote a dynamic, inclusive, and forward-thinking community that bridges the gap between cutting-edge research and practical deployment, bringing together perspectives from both industry and academia.

The goal is to facilitate knowledge exchange, active collaboration, and the celebration of innovation. In this spirit, the group aims to become a key reference point for sustained progress in the field of Generative Edge AI. Each objective reflects the belief that success in this domain depends on the convergence of diverse expertise, from hardware to software, from academic inquiry to real-world engineering.

The following are the Working Group core objectives:

- **Foster Knowledge Sharing**: Facilitate the exchange of ideas and insights through seminars, tutorials, roundtable discussions, and whitepapers.
- **Promote Collaboration**: Build meaningful connections between academia, industry, and individual innovators to drive collective progress in Generative Edge AI.
- **Highlight Achievements**: Recognize and amplify the contributions of members actively shaping the field to inspire and attract new participants.
- **Educate the Community**: Provide accessible resources and updates on the latest breakthroughs, trends, and advancements in Generative Edge AI.

- **Encourage Innovation**: Nurture a culture of exploration and creativity by sharing demos, showcasing individual contributions, and supporting cutting-edge initiatives.

Deliverables

The Working Group is committed to producing tangible outcomes that benefit both the community and the broader AI ecosystem. These deliverables are defined to support learning, promote collaboration, and accelerate the responsible deployment of generative technologies at the edge.

From educational content and hands-on resources to recognition programs and cross-sector publications, the group's outputs are meant to serve as building blocks for continued innovation. In particular, the working group will maintain a strong focus on open access, interoperability, and practical relevance, ensuring that its contributions are both accessible and impactful across the edge AI landscape.

Educational Content

- Tutorials, webinars, and seminars covering both foundational and advanced topics in Generative Edge AI.
- Whitepapers and reports detailing industry trends, research advancements, and best practices.

Community Engagement Activities

- Roundtable discussions to foster dialogue between academia, industry, and individual contributors.
- Networking events to build relationships and encourage collaboration across sectors and disciplines.

Knowledge Dissemination

- Regular updates on breakthroughs, tools, and technologies in Generative Edge AI
- Curated newsletters summarizing key developments and insights from the field.

Recognition and Amplification

- Case studies and success stories showcasing member contributions and achievements.
- Spotlight series on individuals and organizations advancing the field.

Practical Resources

- Demonstrations and walkthroughs of innovative Generative Edge AI solutions.
- Open-access repositories for tools, datasets, and frameworks to enable reproducibility and reuse.

Future-Oriented Initiatives

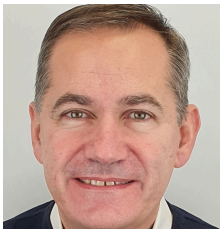
- A dynamic and evolving definition of Edge AI that reflects current advancements in hardware, software, and applications.
- Strategic plans to attract new participants, foster innovation, and ensure the community remains inclusive and forward-looking.

Collaborative Publications

- Co-authored articles, research papers, or blog posts between academic and industry members.
- Annual reviews summarizing the group's impact and the broader progress in the field.

Working Group Leadership

The **Generative Edge AI Working Group** is led by two internationally recognized experts in the field of edge computing and AI:



Danilo Pietro Pau (STMicroelectronics)
Danilo Pau (Fellow, IEEE) received the degree from the Politecnico di Milano in 1992. He joined STMicroelectronics, where he worked on HDMAC and MPEG2 video memory reduction, video coding, embedded graphics, and computer vision. His current work focuses on developing solutions for deep learning tools and applications.

With over 80 patents, 104 publications, 113 MPEG authored documents, and 39 invited talks/seminars at various worldwide universities and conferences, his favorite activity remains mentoring undergraduate students, M.Sc. engineers, and Ph.D. students from various universities. He is currently a member of the IEEE Region 8 Action for Industry and the Machine Learning, Deep Learning and AI in the CE (MDA) Technical Stream Committee IEEE Consumer Electronics Society (CESoc).

Prof. Hajar Mousannif (Cadi Ayyad University)
Hajar Mousannif is a Full Professor at Cadi Ayyad University in Morocco, with over 19 years of experience in Artificial Intelligence, Machine Learning, and Data Science. She has published more than 100 research papers and holds several AI patents. She founded the first Bachelor's and Master's programs in Artificial Intelligence at her university. Hajar also co-chairs the Generative Edge AI Working Group (EDGE AI

FOUNDATION) and the Artificial Intelligence Working Group at the OPCW (Organization for the Prohibition of Chemical Weapons). She is an active member of the global AI community and regularly speaks at conferences to promote responsible and impactful AI development.

Should any other member be added?

Community Momentum: Generative Edge AI on the Edge Forum

Even before the official formation of the Generative Edge AI Working Group, the Edge AI Foundation recognized the transformative potential of generative models at the edge. This vision was brought to life through two editions of the Generative Edge AI on the Edge Forum, which gathered global experts to discuss cutting-edge research, share practical insights, and explore future directions for generative intelligence in resource-constrained environments.

In March and October 2024, the first two forums became cornerstone events, marking the transition from TinyML to a broader conversation around Generative Edge AI. They laid the groundwork for the working group's creation and remain a core part of its ongoing activities, showcasing the community's commitment to open dialogue, interdisciplinary collaboration, and real-world impact.

Since then, a surge of innovation has followed, new studies, novel applications, and a better understanding of edge-specific use cases. The EDGE AI FOUNDATION community continues to express a strong desire to stay up-to-date, share knowledge, and build a common foundation for the future of generative edge intelligence.

You can revisit the presentations from both editions here:

- March 2024 – Generative Edge AI on the Edge Forum (YouTube Playlist)
- October 2024 – Generative Edge AI on the Edge Forum (YouTube Playlist)

The journey continues with the third edition of the *Generative Edge AI on the Edge Forum* (link), a two-day livestream event focused on the impact of Generative Edge AI platforms, highlighting progress in hardware, software, tooling, applications, and services, and exploring emerging paradigms such as agentic and physical AI.

Stay tuned to the Generative Edge AI Working Group page for updates, recordings, and opportunities to participate in upcoming events.

Highlights from the First Generative Edge AI on the Edge Forum

The inaugural *Generative Edge AI on the Edge Forum* set the stage for a vibrant, interdisciplinary exchange around deploying generative models on resource-constrained platforms. With contributions from academia, industry, and the open-source community, the event covered both visionary ideas and hands-on engineering advances. Key themes included:

- **Miniaturized LLMs and Efficient Inference**
  Talks by Syntiant, NXP, and Arm highlighted strategies for distilling and quantizing LLMs to run efficiently on embedded platforms, including the use of NPUs, custom SoCs, and advanced model optimization techniques.
- **Generative AI for Hardware Design**
  Speakers from Harvard, UC Davis, and Efabless explored how foundation models can be used to accelerate chip design, optimize architectures, and even auto-generate Verilog for edge-specific hardware
- **Edge Applications in Real-World Domains**
  Sessions from Bosch, Qualcomm, UNICEF, and Johns Hopkins showcased how Generative Edge AI is being applied to domains such as connected vehicles, education, healthcare, and embodied systems—often leveraging novel data modalities and hybrid architectures.
- **Human-AI Interaction and Design Futures**
  Contributions from IDEO and Useful Sensors pushed the boundaries of how Generative Edge AI systems should interact with humans, with alternative models of AI experience inspired by calm technology and creative narratives
- **Research Frontiers and System-Level Thinking**
  Presentations by EPFL, Meta, and others offered a forward-looking lens on emerging capabilities—such as multimodal foundation models, agentic AI, and strategies for lifelong learning and adaptation at the edge.

Highlights from the Second Generative Edge AI on the Edge Forum

Building on the momentum of the first event, the second edition of the *Generative Edge AI on the Edge Forum* continued to expand the community's understanding of deploying generative models in edge environments. The forum featured leaders from academia, industry, and research institutes, offering a wide-angle view of current innovations and real-world challenges.

Key highlights included:

- **Edge Infrastructure & Strategic Perspectives**
  Dave McCarthy of IDC opened the forum with a forward-looking perspective on how LLMs and transformer models are reshaping the edge computing landscape, accelerating adoption and infrastructure readiness.

- **Model Deployment & Optimization**
  Talks from Meta, Arm, and ETH Zurich explored techniques for compressing and optimizing generative models to fit within the tight constraints of edge hardware, including use of ExecuTorch, RISC-V SoCs, and ARM MPUs.
- **Lifecycle Integration & TinyML Synergies**
  EURECOM and Fondazione Bruno Kessler presented work on merging TinyML lifecycles with LLMs and deploying advanced generative applications—such as neural style transfer—on ultra-low-power MCUs.
- **Domain-Specific Applications**
  BOSCH and Wipro shared lessons from deploying Small Language Models in automotive and enterprise contexts, with applications ranging from custom code generation to in-vehicle personalization.
- **New Approaches to Privacy, Memory & Security**
  Speakers from NXP, Kyung Hee University, and the Technology Innovation Institute discussed advances in memory optimization, secure fine-tuning, and model compression, using examples like Falcon Mamba and privacy-preserving inference.
- **Tools, Platforms & Future Directions**
  The forum also showcased community-driven tools such as TinyRAG, hardware design strategies like SECDA-LLM, and deployment considerations for 5G edge platforms shared by Particle.io.

This second forum reinforced the community's shared belief that Generative Edge AI at the edge is not just possible—it's already happening, and it requires continued collaboration across disciplines to scale responsibly, efficiently, and inclusively.

Help Shape the Future: Generative Edge AI Survey

As part of its commitment to inclusive innovation and global collaboration, the Generative Edge AI Working Group has launched a strategic community survey. Initially shared with partners of the Edge AI Foundation, this survey aims to gather insights from key stakeholders across academia, industry, and the open-source ecosystem to inform the group's priorities, initiatives, and outputs.

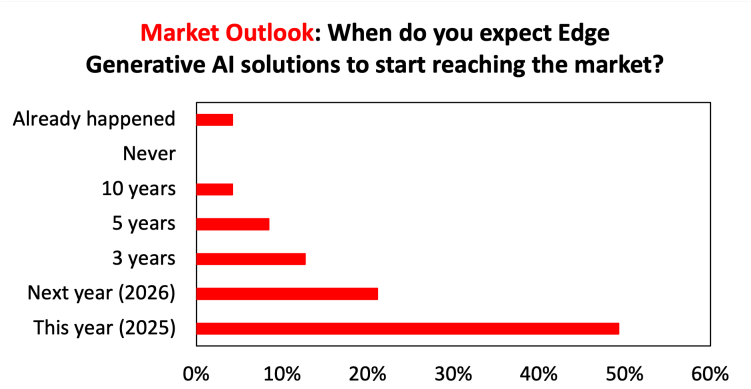
The questionnaire explores a wide range of topics, from technical readiness and adoption barriers to preferred application domains, collaboration formats, and emerging trends. It also captures early community sentiment on key topics such as Agentic AI at the edge, education and outreach needs, and the types of deliverables that would bring the most value to participants.

Here's a brief summary of initial findings and trends, which reflect early community input:

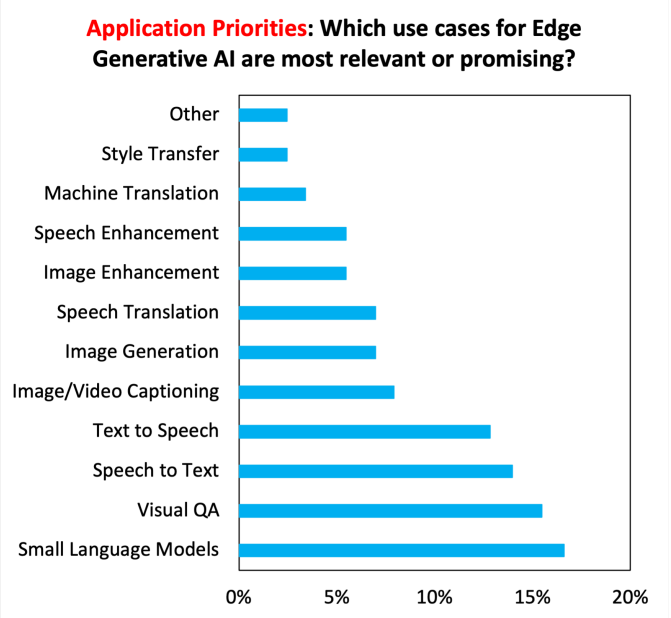
Survey Highlights

The initial wave of responses from the Generative Edge AI Working Group community survey offers a timely snapshot of expectations, priorities, and barriers in the evolving Generative Edge AI landscape.

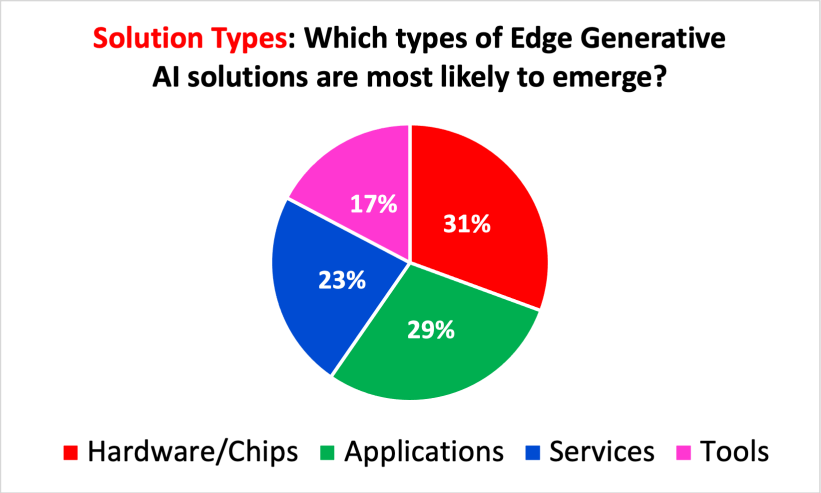
Market timing expectations are optimistic: a clear majority of respondents (over 70%) anticipate that Generative Edge AI solutions will begin appearing on the **already in 2025**, with significant momentum expected to continue into 2026 and beyond. Only a small fraction projected timelines beyond 5 years or expressed uncertainty.



When asked about **preferred applications**, the community showed strong interest in **Small Language Models**, **Visual Question Answering**, **Speech-to-Text**, **and Text-to-Speech** technologies. These were followed by media-based use cases such as captioning, generation, and enhancement—underscoring the perceived value of multimodal generative capabilities in constrained environments.



On the **solution front**, respondents expect to see impact across the stack: **hardware/chips**, **applications**, and **services** were the most anticipated areas, with **tools** also seen as important enablers.



Beyond technical priorities and adoption timelines, the survey revealed several important trends shaping the direction of Generative Edge AI.
Adoption is primarily driven by the desire to improve human-machine interaction and to enable novel AI-native products, both cited by over 76% of respondents. Closely behind, over 70% highlighted the emergence of use cases that were previously not possible with traditional AI approaches.

In terms of **organizational focus**, product development and R&D lead the way, with 88% and 76% of respondents prioritizing them, respectively. Model deployment, while still relevant, was seen as secondary, suggesting that the community is still in a foundational exploration phase.

Collaboration interests reflect how organizations wish to engage with others in the ecosystem. The most common preference was for **use-case–driven projects** (82.4%), followed by **collaborations around datasets and customer initiatives** (64.7%), and **joint research efforts or technical workshops** (58.8%). These responses point to a strong desire for partnerships that are grounded in practical relevance and mutual experimentation, rather than abstract or siloed efforts.

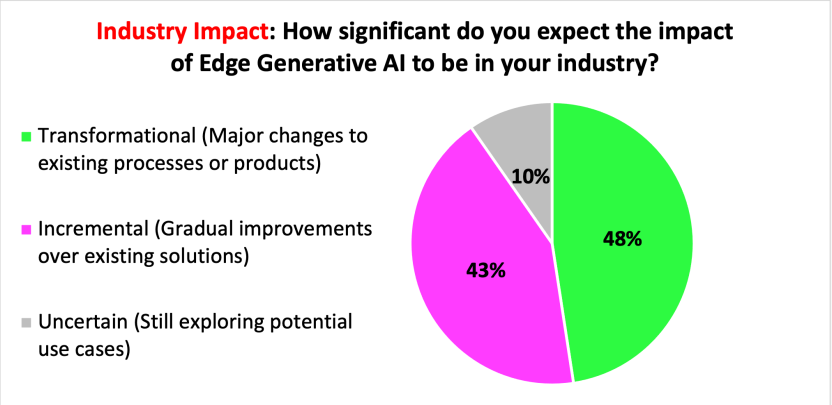
When asked about **desired forms of support from the foundation and the working group**, the top responses included open-source initiatives, real-world case studies and demos, and access to cutting-edge research. In contrast, areas like policy guidance and access to large-scale compute resources were noted as lower priority for many respondents at this stage.

Several **emerging trends** were also identified. *IoT and Industrial applications* topped the list of sectors to watch, followed by *consumer-facing systems*, *humanoid robotics*, and *multimodal AI*.

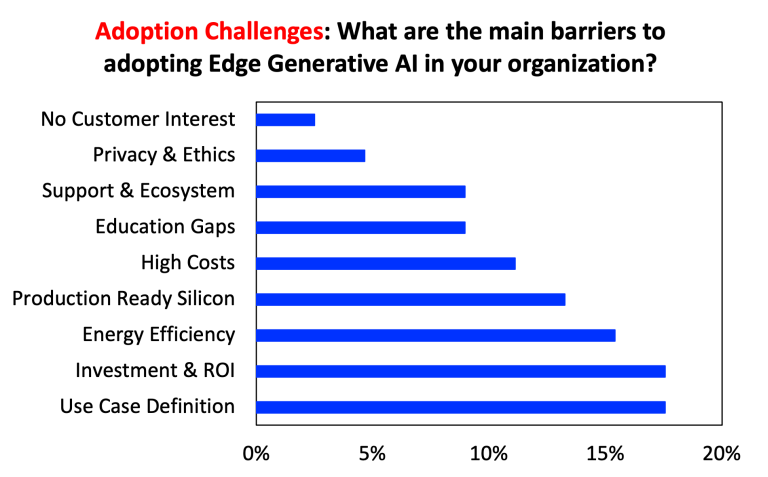
The community also showed strong interest in **Agentic AI at the edge**, with over 76% supporting further exploration of the topic. That interest, however, was often paired with concerns about safety, hallucination risks, and trustworthiness—suggesting a need for transparent frameworks and continued education.

Multiple comments emphasized the need for **proof-of-concept deployments** and **educational content**, especially around new paradigms like agentic and autonomous systems. While excitement is clearly growing, practical grounding and responsible innovation remain top of mind.

The **perceived impact** of Edge Generative AI is overwhelmingly positive, with nearly all respondents rating it as either transformational or incremental, and very few expressing uncertainty or skepticism.



Finally, the survey highlighted key **adoption barriers**, led by the **definition of use cases**, **ROI/investment concerns**, and energy efficiency limitations. The lack of **production-ready silicon**, **high implementation costs**, and **education gaps** were also cited frequently, suggesting where coordinated action and resources could have the most immediate effect.



These insights are helping to inform the working group's agenda and will guide future initiatives.

We are considering opening the survey to the broader public. Whether your organization is already active in Generative Edge AI or just beginning to explore its potential, your input can help steer the group's direction and ensure that its work is aligned with real-world challenges and opportunities.

Your voice matters, and together, we can build a stronger, more connected, and impactful Generative Edge AI ecosystem.

Stay tuned for updates and future opportunities to contribute!!

Get Involved

Whether you're developing models, building systems, optimizing hardware, or exploring novel applications, **the Generative Edge AI Working Group welcomes your voice**. Join us in shaping a future where generative intelligence is accessible, efficient, and embedded at the very edge of our connected world.