Expanding The Horizons of Generative Edge AI: Mission, Vision, and Insights From Industries

Roberto Morabito
Communication System Department
EURECOM
Valbonne, France
roberto.morabito@eurecom.fr

Riccardo Adorante
System Research and Applications
STMicroelectronics
Agrate, Italy
riccardo.adorante@st.com

Hajar Mousannif
Department of computer science
Cadi Ayyad University
Marrakesh, Morocco
hajar.mousannif@gmail.com

Danilo Pietro Pau, FIEEE
System Research and Applications
STMicroelectronics
Agrate, Italy
danilo.pau@st.com

***Abstract*—The Generative EDGE AI Working Group, established since 2024 within the EDGE AI FOUNDATION, is dedicated to advancing the responsible adoption of generative artificial intelligence (AI) at the edge and on-premises. This paper provide a report of its operations outlines the group's mission to connect academia, industry, and the open-source community through collaborative research, educational initiatives, and community-driven activities. It presents a vision for deploying generative models, such as small language models (SLMs), multimodal AI, and agentic systems, for resource-constrained devices. In addition to detailing the group's objectives and deliverables, this work includes a comprehensive summary of the speeches delivered by leading industries and research entities during the third Forum of the GenAI on the Edge series. These summaries highlight both qualitative and quantitative insights shared by global experts, covering topics such as model optimization, hardware innovations, real-world applications, and emerging paradigms like agentic AI. By integrating these perspectives, the paper provides a holistic view of the current state and future potential of generative AI at the edge. This effort aims to establish a foundation for scalable, inclusive, and impactful innovation in generative intelligence, positioning the working group as a key reference point for the community driving the transformation from cloud centric to both on premises and edge centric edge AI**

***Index Terms*—EDGEAI Foundation; artificial intelligence; edge; on premises; generative; small language models, tiny machine learning**

I. Introduction

The Generative EDGE AI (GenEdgeAI) Working Group (WG) is a collaborative initiative within the EDGE AI FOUNDATION dedicated to the advancement of the frontiers of generative AI to achieve in real-time, resource-constrained deploymemts, and decentralized (edge and on premises type of systems) environments.

Large-scale generative models continue to reshape how people interact, through the cloud, through technology spanning from multimodal assistants and real-time translation to autonomous systems and industrial monitoring. These bring generative capabilities to both the edge and on premises as the next bold step in AI democratization with an expected impact to everyone in everyday life.

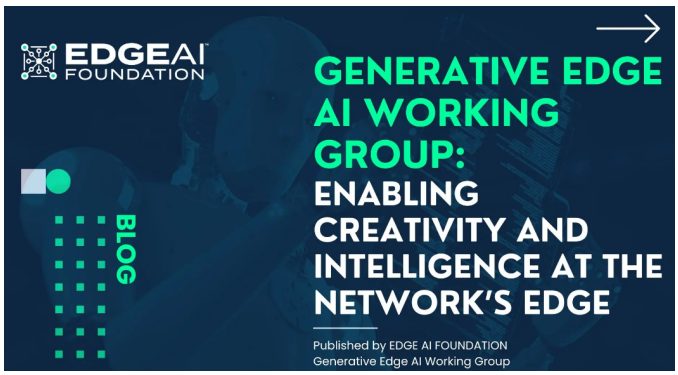

Fig. 1: Working Group Landing page. Courtesy by EdgeAI Foundation

This WG brings together academic researchers, industry experts, and open-source contributors to make this vision a reality.

The WG aim to provide and share convincing arguments such that the industries will reach a point to invest on product and services making Edge GenAI a reality as TinyML demonstrated to achieve with Fixed AI in the past years. This WG will empower edge and on premises devices with generative AI capabilities that are energy-efficient, privacy-preserving, data-sovereignty, responsive, and autonomous, unlocking intelligent behavior closer to the local premises, the user, the data generationvia sensors, and by providing to the user with natural machine interaction experience.

Charter of the Working Group

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, software

and service solutions capable of enabling natural interaction on edge devices at extremely high energy efficiency levels, typically in the tera to peta operations per Watt (TOPS/W to POPS/W) range.

This new field is expected to enable an unprecedented generation of powerful yet energy efficient neural processor units (NPUs), in-memory computing, and their integration into systems-on-chip (SoCs) that leverage heterogeneous integration to support scalable and sustainable edge intelligence at an affordable cost to everyone.

Definition of the Working Group

Generative EDGE AI refers to deploying and running generative AI models directly on edge devices (e.g., smartphones, internet of things (IoT) devices, sensors, micro controllers, multi processors, autonomous devices) rather than relying on centralized and re-motized 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 of the Working Group

The Generative EDGE AI WG 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 of the Working Group

To fulfill its mission, the Generative EDGE AI WG 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 WG core objectives:

- Foster Knowledge Sharing: Facilitate the exchange of ideas and insights through seminars, tutorials, round-table discussions, and white-papers.
- 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 of the Working Group

The WG is committed to produce 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 WG will maintain a strong focus on open access (such as this paper), interoperability, and practical relevance, ensuring that its contributions are both accessible and positively impacting the edge AI landscape.

Educational Content

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

Community Engagement Activities

- Round-table 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 walkthrough 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, tools and applications.
- Strategic plans to attract new participants, foster innovation, and ensure the community remains inclusive and forward-looking.

Collaborative Publications

- Co-authored articles (such as this manuscript), 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 Pau (STMicroelectronics) is Technical Director in System Research and Applications. He is an engineer and researcher in the field of AI and machine learning (ML) at the edge. He has contributed to numerous innovative projects about multimedia processing (video, graphics), computer vision and artificial intelligence. He is active since 1991 following his studies at Politecnico di Milano. Nowadays, his work primarily focuses on the development and application of AI technologies (neural architecture search, high parameter optimizations, on device learning, generative AI in various application domains, including embedded systems and edge computing. Danilo Pau has authored and co-authored several research papers and patents, showcasing his expertise and contributions to the field. He is known for his deep understanding of AI algorithms and their practical implementations, which have had a significant impact on the industry. In addition to his technical achievements, Danilo Pau is also recognized for his efforts in promoting AI education and collaboration within the tech community, including IEEE and EdgeAI Foundation in which he serves in different ways. He often participates in conferences, workshops, and seminars, sharing his knowledge and insights with peers and aspiring engineers.

Prof. Hajar Mousannif (Cadi Ayyad University) is a Full Professor at Cadi Ayyad University in Morocco, with over 19 years of experience in AI, ML, 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 AI at her university. Hajar also co-chairs the Edge Generative AI Working Group (Edge AI Foundation) and the AI 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.

Community Momentum: GenAI at the Edge Forum

It shall be recognized that, even before the official creation of the Generative EDGE AI WG, the Edge AI Foundation envisioned the transformative potential of generative models at the edge. This vision was brought to life through two editions of the GenAI 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 need to stay up-to-date, share knowledge, and build a common foundation for the future of generative edge intelligence.

Any interested expert in the field, can revisit the presentations from both editions here:

- March 2024 – GenAI on the Edge Forum (YouTube Playlist, last access 2025/7/10)
- October 2024 – GenAI on the Edge Forum (YouTube Playlist, last access 2025/7/10)

This journey continued with the third edition of the GenAI on the Edge Forum, 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.

- Day One, May 2025 – GenAI on the Edge Forum (YouTube Playlist, last access 2025/7/10)
- Day Two, May 2025 – GenAI on the Edge Forum (YouTube Playlist, last access 2025/7/10)

The GenAI at the Edge WG webpage, (last access 2025/7/10) reports updates, recordings, and opportunities to participate in upcoming events.

Highlights from the First GenAI on the Edge Forum

The inaugural GenAI 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 University showcased how GenAI 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

GenAI 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 GenAI on the Edge Forum

Building on the momentum of the first event, the second edition of the GenAI 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 (FBK) presented work on merging TinyML lifecycles with LLMs and deploying advanced generative applications—such as neural style transfer—on ultra-low-power MCUs to achieve content anonymization.

Domain-Specific Applications. BOSCH and Wipro shared lessons from deploying SLMs 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, last access 2025/7/10) , hardware design strategies like SECDA-LLM, last access 2025/7/10), and deployment considerations for 5G edge platforms shared by Particle.io.

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

Insights from Industry Experts: Summaries of the Third Edge AI Foundation Forum

The third edition of the Edge AI Foundation Forum, held online on May 27 and 28, 2025, brought together leading voices from renown academia researchers and industry leaders to discuss the latest advancements in generative AI at the edge. Over two days, experts delivered insightful speeches that explored both qualitative and quantitative dimensions of generative AI, covering topics such as hardware innovations, model optimization, real-world applications, and emerging paradigms like agentic AI. This section provides a comprehensive summary of these speeches, highlighting key takeaways and actionable insights. By gathering all the diverse perspectives and experiences shared during the forum, this section aims to offer to the readers a deeper understanding of the current state and future potential of generative and Agentic AI on the Edge. Let's deepen them one by one.

Fireside chat with Marketing Analyst

The transformation of the network paradigm from the "Internet of Everything" to the "Intelligence of Everything" marks a pivotal shift in the integration of AI into edge computing. This transition, as highlighted by the speaker, underscores the movement of native AI from distant cloud services to the edge, enabling faster, more localized decision-making. The convergence of GenAI with edge AI solutions is poised to redefine industries, with commercial solutions anticipated by 2026.

The speaker, a seasoned marketing analyst with over five years of experience in AI, emphasized the rapid evolution of edge AI. He framed this shift using insights from researchers, noting the burst of edge AI advancements over the past decade, driven by hardware innovations from companies like NVIDIA. A standalone report on generative edge AI, published in 2024 [Soldatos and Rao(2024)], highlights the convergence of generative models with edge hardware, showcasing the potential for transformative applications.

The CEO of the EdgeAI Foundation, provided a complementary perspective. He detailed the evolution of generative edge AI demos, which began in March 2024, and have since then become faster and more responsive. He also mapped the complexity of edge deployments across "far edge", "near edge", and hybrid cloud–edge scenarios, identifying industrial automation as the most promising early adopter. However, he cautioned against over hyping the technology, citing regulatory barriers and workforce skill gaps as significant challenges.

The timeline for edge AI's growth is compelling. The market analyst noted the acceleration of edge-related hardware demos over the past decade, culminating in the release of his 2024 report. The CEO forecasted an 18-month window for transitioning from prototypes to commercial deployment, targeting late 2025 or early 2026. He also outlined three primary research areas for exploration:

- Augmenting edge AI with vision–language models.
- Developing human–machine interfaces using natural language.
- Integrating edge technology into agentic AI frameworks.

Both speakers identified industrial automation as the leading sector for early adoption, with retail and entertainment as fast followers. Wearables, smartphones, and PCs were highlighted

as economic engines for edge AI, while human–machine conversational interfaces offer unique qualitative benefits. However, regulatory constraints in industries like automotive and aerospace, along with workforce skill gaps, remain barriers to widespread adoption.

The maturation of edge-based generative AI is evident, but its true commercial impact will likely crystallize over the next 12–24 months. Key areas of interest include:

- Vision–language integration on edge sensors.
- Agentic protocols for interoperable AI endpoints.
- Physical AI in robotics, leveraging foundational models for observational learning.

As the field progresses, stakeholders must address regulatory and workforce challenges to unlock the full potential of generative AI in edge computing.

Building Optimized Gen AI Use Cases with OpenVINO

The evolution of human-computer interaction has transitioned from traditional input methods, such as keyboards and mice, to more intuitive and natural interfaces. GenAI is at the forefront of this transformation, enabling devices to act as personal assistants that enhance daily life. This discussion explores the qualitative and quantitative advancements in conversational AI, optimization techniques, and the integration of software-hardware ecosystems.

The advent of chatbot technology has redefined how users interact with computers. These conversational interfaces allow for natural language communication, addressing real-world challenges such as hospital check-ins. For instance, chatbots can guide patients, reducing waiting times and improving their overall experience. The vision extends to localized generative AI assistants embedded in devices like PCs, smartphones, and wearables. These assistants can perform diverse tasks, from monitoring a baby's activities to assisting with paint selection using voice and vision inputs.

The question of "where to put the compute" is critical in enabling efficient AI on edge devices. By leveraging optimization techniques, such as post-training quantization (PTQ), weight compression, and filter pruning, AI models can be significantly reduced in size and computational demand while maintaining performance. For example, the OpenPose model was compressed from 16.6 MB to 4.7 MB, achieving a near-doubling of frame rates (42.5 FPS to 90 FPS) with minimal performance degradation. This optimization is particularly vital for large language models (LLMs) on edge devices, where reducing RAM usage by compressing weights from 32 bits to 8 bits results in a fourfold reduction in size with only a 5% increase in perplexity.

Intel's Open Edge Platform exemplifies the integration of software and hardware to streamline AI deployment. This ecosystem spans CPUs, integrated GPUs, discrete GPUs, and integrated NPUs, unified through open-source tools like OpenVINO. Developers benefit from automated quantization and cross-platform deployment, reducing the development cycle for edge AI demos from months to just two weeks. The platform also provides extensive resources, including 200 Jupyter notebooks and 20 reference design kits, enabling seamless model optimization and deployment. Table I highlights key quantitative improvements enabled by optimization and hardware integration. The possibility to integrate GenAI into personal devices fosters a future where on-device intelligence becomes the norm. Native chatbot experiences can now operate without sending sensitive data to the cloud, ensuring privacy and security. Optimization techniques, such as automated quantization, unlock significant performance gains while reducing memory and power requirements. Developers are encouraged to explore platforms like Intel's Open Edge Platform to experience real-time transcription, image generation, and lightweight LLM chatbots. In conclusion, the combination of conversational AI, model optimization, and integrated software-hardware ecosystems is revolutionizing human-computer interaction. These advancements not only enhance user experiences but also pave the way for more efficient and accessible AI solutions.

Generative AI at NXP – Bringing intelligence to the Edge

The presentation highlighted the transformative potential of GenAI in edge computing, emphasizing the company's "Edge First AI Strategy." This approach prioritizes inference at the edge over cloud-based model training, ensuring private, secure, and efficient AI solutions for industrial automation, smart homes, automotive systems, and more. By leveraging advanced discrete NPUs and modular software pipelines, NXP is redefining the boundaries of edge AI deployment.

NXP distinguishes between AI development, which involves training models on cloud or local servers, and AI deployment, which focuses on enabling edge devices to execute AI tasks. Deployment is particularly critical in fields such as industrial automation, building and energy management, and automotive systems like in-cabin monitoring and advanced driver-assistance systems (ADAS). The introduction of multimodal AI solutions and increasingly powerful discrete NPUs has expanded the scope of edge AI applications, enabling innovative use cases like smart medical assistants capable of visual patient screenings.

The speech presented the eIQ GenAI Flow: a comprehensive software pipeline designed by NXP to fine-tune and optimize LLMs for edge deployment. This pipeline adapts open-source LLMs to domain-specific data, reducing hallucinations and errors while safeguarding sensitive information. Fine-tuning is particularly essential for applications in medical diagnostics, automotive systems, smart buildings, and factory automation. Additionally, the pipeline supports quantization, shrinking memory footprints to as low as 4 bits while maintaining accuracy, and deploying models efficiently on NXP's NPUs.

An example of the eIQ GenAI audio processing flow demonstrates the pipeline's capabilities:

- Input: Audio signal
- Processing Steps: Wakeword engine $\rightarrow$ Automatic speech recognition (ASR) $\rightarrow$ LLM with retrieval-augmented generation (RAG) $\rightarrow$ Text-to-speech (TTS)
- Output: Natural-sounding speech

TABLE I: Key quantitative improvements enabled by optimization and hardware integration

Metric / Use Case	Before Optimization	After Optimization
Teddy Bear Image Generation	8 s on an integrated GPU	2 s on a discrete Arc GPU (16 GB RAM)
Llama 7B Memory Footprint	$\sim$25 GB RAM at 32-bit	$\leq$8 GB after 4-bit quantization
Movidius USB on Raspberry Pi	CPU-only Pi performance	10$\times$ speedup at just 5 W power draw
OpenPose Model Size	16.6 MB	4.7 MB
OpenPose Inference Time	23.6 ms (42.5 FPS)	11.2 ms (90 FPS)

The integration of RAG enhances the model's knowledge by enabling the creation of compatible databases, further improving the pipeline's utility.

In collaboration with Kinara, NXP showcased a multimodal assistant running entirely on the i.MX 8M Plus "FRDM" board with a Kinara ARA-2 NPU. This assistant combines text, vision, and speech capabilities, exemplified by the LLaVA (vision-language assistant) model. The system processes images and generates spoken responses without cloud connectivity, demonstrating private, low-latency smart camera applications. Table II summarizes the characteristics and performance metrics achieved by the different components of the NXP's GenAI pipeline.

In conclusion, NXP's advancements in GenAI for edge computing underscore the company's commitment to deliver secure, efficient, and scalable AI solutions. By focusing on private, on-device intelligence and leveraging cutting-edge hardware and software innovations, NXP is paving the way for transformative applications across industries.

Industry Applications and deployment

The integration of multimodal artificial intelligence (MAI) into traditionally passive devices, such as printers and scanners, represents a transformative opportunity to enhance their functionality. Wipro Limited, a global leader in IT consulting and software services, is spearheading efforts to embed multimodal AI into these devices, enabling them to interpret, reformat, and generate content autonomously. This discussion highlights the qualitative and quantitative aspects of this initiative, focusing on its applications, deployment strategies, and challenges.

Printers and scanners, which have historically been limited to passive operations, are now being reimagined as intelligent endpoints. By leveraging MAI, these devices can process complex inputs, such as multilingual documents, and produce structured outputs without relying on cloud-based solutions. This shift not only enhances user experience but also addresses privacy concerns by enabling on-device processing.

Wipro Limited imagined two key use cases for GenAI applications on the Edge:

- Document Understanding and Reformatting: Multimodal AI models, such as LayoutLMv3 and TATR, can extract tabular data from scanned images and reformat it into graphs or charts. This capability is particularly useful for creating accessible layouts for visually impaired users or removing extraneous elements like advertisements.
- Creative Media Generation: Generative diffusion models, such as Flux, enable the creation and iterative modification of images. For instance, users can generate personalized greeting cards by providing simple text prompts.

Regarding the End-to-End data processing pipeline for these applications, the following steps have been individuated by Wipro:

- Input Acquisition: Scanned PDFs or images, optionally accompanied by voice or text prompts.
- Preprocessing: Tasks such as resolution downscaling, denoising, and page segmentation prepare the input for analysis.
- Core Inference: Visual language models (VLMs) like QWEN 2.5 VL handle layout, text, and image comprehension, while generative diffusion engines synthesize or modify images.
- Postprocessing: Outputs are cleaned and reformatted, such as converting tables into charts or applying new designs to image masks.
- Deployment Options: These include fully on-device processing, offloading to companion PCs or AI accelerators, and hybrid cloud solutions.

Despite the potential, deploying multimodal AI in edge devices like printers presents significant challenges:

- Resource Constraints: Modern VLMs and diffusion models require substantial compute power and memory, with some models demanding up to 24 GB of VRAM.
- Edge-Ready Strategies: Techniques such as model quantization, hyperparameter tuning, and the use of GPU-optimized runtimes (e.g., NVIDIA TensorRT) help reduce the computational footprint. Companion devices, such as AI-enabled PCs or smartphones, can also host heavy models, with printers serving as input/output endpoints.
- Future Roadmap: As SoC designs evolve to include more NPUs and larger memory capacities, fully embedded AI solutions in printers will become feasible.

Table III summarizes the details and requirements of the VLM and Generative diffusion pipelines.

Wipro's exploration of MAI in printers and scanners demonstrates the feasibility of deploying advanced AI capabilities at the edge. By combining techniques such as model quantization, hyperparameter tuning, and hybrid edge architectures,

TABLE II: The following table summarizes the performance metrics of key components in NXP's GenAI pipeline

Component	Model & Size	Quantization	Hardware	Performances
ASR	Whisper small (244M params)	8-bit	i.MX 95 Neutron NPU (2 TOPS)	Processes 3-second audio segments in real time for speech-to-text conversion.
LLM (Chat)	H2O Danube (500M params)	8-bit	Neutron NPU (2 TOPS)	$\sim$0.5 seconds to first token; low CPU usage (24%) due to offloading.
TTS	Bits (19.5M params)	8-bit	Neutron NPU	Produces natural-sounding speech with multi-voice support.
Multimodal Assistant	LLaVA: LLaMA 3 8B + CLIP 428M	4-bit/8-bit	Kinara ARA-2 NPU (40 eTOPS, 16 GB)	CLIP: 430 ms per image; LLaMA 3 8B: 6.5 tokens/s; TTFT (new image): 7 s.

TABLE III: Summary of key aspects and details of the VLM and diffusion model pipeline

Aspect	Details
VLM Parameterizations	QWEN 2.5 VL operates with 3 billion or 7 billion parameters, quantized to 4 bits.
VLM Memory Footprint	Requires 16 GB of VRAM for the 7 billion parameter model.
Diffusion Model (Flux)	A hybrid multimodal architecture with 12 billion parameters, requiring 24 GB of VRAM.
Hardware Target	Proof of concept tested on x86 servers with NVIDIA T4 GPUs (16 GB VRAM) using TensorRT.
Quantization Efficiency	Four-bit quantization reduces model size by approximately 4$\times$, enabling deployment on commodity GPUs.
Pipeline Optimization	Image downscaling (e.g., 4K to 1K) accelerates preprocessing by $\sim$4$\times$ with minimal quality loss.

enterprises can unlock new use cases today. As hardware capabilities continue to advance, fully embedded AI solutions in printers and scanners are poised to become a reality, revolutionizing document workflows and creative media generation.

When is GenAI Useful at the Edge?

GenAI has traditionally been associated with cloud-based systems due to its computational and memory-intensive nature. However, the speaker's insights highlight the untapped potential of deploying GenAI at the edge, leveraging the vast compute capacity of embedded devices such as microcontrollers (MCUs). This discussion explores the qualitative and quantitative aspects of GenAI at the edge, focusing on its challenges, optimization strategies, and practical use cases.

The edge offers unique advantages for GenAI workloads, including reduced latency, enhanced privacy, and offline operation. It is estimated that the combined compute capacity of all MCUs worldwide is approximately twice that of all GPUs, underscoring the potential of edge devices. These benefits make the edge a compelling environment for deploying GenAI, particularly in scenarios where real-time responses and data privacy are critical.

Transformers, the backbone of modern GenAI models, present significant challenges for edge deployment. Their memory-bound nature, driven by the attention mechanism, necessitates the use of DRAM, which increases power consumption to hundreds of milliwatts—unsuitable for low-power systems. Additionally, the large weight matrices required for each token exacerbate memory limitations. To address these challenges, several optimization strategies have been proposed:

- Tiling and Blocking: Reducing on-chip memory requirements by generating activations in smaller, manageable blocks.
- Specialized DSPs: Leveraging digital signal processors (DSPs) like Qualcomm's Hexagon, which offer advanced memory handling capabilities.
- Cache Optimization: Utilizing tightly coupled memory for efficient prefetching of weight blocks, thereby improving performance predictability.

Despite the challenges, several practical applications of GenAI at the edge are emerging:

- Knowledge Retrieval: Local LLMs can answer up to 80% of simple search queries, making them ideal for static corpora such as device manuals or store FAQs. This eliminates the need for cloud connectivity, enhancing privacy and reducing latency.
- Voice Interfaces and Actions: GenAI can enable advanced speech-to-intent pipelines, allowing users to control physical devices through natural language commands. However, grounding these systems in real-world contexts remains a challenge.
- Voice Synthesis: On-device text-to-speech systems, including voice cloning, are nearing maturity. These systems promise natural, personalized voice outputs and could soon become standard features in edge platforms.

Table IV gathers insights shared by the speaker on key GenAI metrics.

The most potential was found for those use cases that emphasize static knowledge bases (e.g., device manuals) or intent-to-action pipelines (e.g., voice commands controlling actuators), where accuracy, determinism, and privacy are paramount. In conclusion, while GenAI at the edge faces significant technical hurdles, its potential for low-latency, privacy-preserving applications is undeniable. By focusing on memory optimization and targeted use cases, the edge can become a viable platform for deploying generative AI.

The Move of AI's Center of Gravity to Edge Devices

The paradigm of AI has undergone a significant transformation, particularly with the advent of GenAI and its integration into edge devices. Historically, AI on the edge was predominantly focused on perception tasks such as detection and classification. However, the last 18 months have witnessed a shift towards generative and agentic AI, enabling more complex and personalized functionalities across various domains, including automotive, personal computing (PC), extended reality (XR), industrial applications, and networking.

Generative AI has introduced capabilities that extend beyond traditional perception tasks. For instance, in the automotive sector, GenAI can assist drivers by diagnosing system failures, linking issues to specific components, and even scheduling appointments by accessing personal calendars. This transition is not limited to automotive applications; it spans text creation, content generation, photo and video editing, live translation, and AI assistants. The next frontier lies in agentic AI, which interprets user intent and orchestrates multi-step actions, such as planning trips or managing schedules, with the potential to execute these tasks either locally or via cloud integration.

Scaling AI across multiple devices is a critical challenge, and Qualcomm has addressed this with its AI stack, comprising three distinct hardware blocks:

- Oryon CPU: Designed for immediate processing needs.
- Adreno GPU: Optimized for image and video processing.
- Hexagon NPU: Tailored for sustained and pervasive AI applications.

This architecture ensures best-in-class performance and power efficiency, enabling on-demand and sustained AI functionalities. Qualcomm's unified AI software ecosystem further supports scalability, integrating frameworks like TensorFlow, PyTorch, and Keras, along with developer tools such as profilers, debuggers, and system software.

The principle of "What happens on the edge stays on the edge" underscores the importance of privacy, reduced latency, and offline operation. Edge devices build personalized knowledge graphs by sharing context across form factors, such as phones, PCs, and AR glasses, enabling seamless multi-device experiences. This approach minimizes reliance on cloud-based solutions, which are increasingly costly for inference tasks, and ensures that sensitive data remains local.

The development of AI models for edge devices has seen a focus on optimizing model sizes to balance performance and resource constraints. For example:

- Handsets typically utilize models with 1–10 billion parameters.
- PCs employ models with 13–20 billion parameters.
- AR/VR glasses operate with smaller models of 1–4 billion parameters.

Aggressive quantization and fine-tuning techniques are employed to reduce memory and compute requirements while maintaining accuracy. Hybrid AI orchestration further optimizes cost and performance by escalating tasks from the device to edge servers and, only when necessary, to the cloud.

Edge-centric AI has evolved to support not only perception tasks but also generative and agentic workloads. Personalized, multi-device knowledge graphs enhance user experiences without compromising privacy. Qualcomm's hybrid orchestration strategy smartly balances privacy, latency, and cost, while its AI stack and ecosystem partnerships enable rapid deployment across industries.

For developers, the next steps include:

1) Exploring the Qualcomm AI Hub to test and deploy models on virtual devices.
2) Prototyping agentic AI flows by combining on-device large language models (LLMs) with voice or text inputs.
3) Architecting applications for hybrid failover, ensuring seamless escalation from device to edge to cloud.

Table V summarizes the principal components of Qualcomm's AI stack.

In conclusion, the shift towards edge-centric AI represents a significant milestone in the evolution of artificial intelligence, offering enhanced privacy, reduced latency, and cost efficiency. By leveraging Qualcomm's AI stack and hybrid orchestration, developers can unlock the full potential of generative and agentic AI across diverse applications.

Optimizing LLMs for Hardware Accelerators

The optimization of LLMs for hardware accelerators is a critical area of focus due to the significant challenges posed by their size and computational demands. LLMs, with their gigabyte-scale memory footprints, require innovative strategies to ensure efficient deployment on devices with limited resources. The speech explores the motivations, challenges, and core optimizations for adapting LLMs to hardware accelerators, as well as the implications for edge computing and future trends.

LLMs are inherently resource-intensive, necessitating optimization to address the following goals:

- Efficiency: Reducing energy consumption to extend battery life in mobile and edge devices.
- Low Latency: Enabling real-time, interactive responses for applications requiring immediate feedback.
- Cost Reduction: Minimizing the expenses associated with cloud-based inference by shifting computation to on-device processing.

TABLE IV: Summary of key metrics and observations related to Edge MCU and LLM performance

Metric / Aspect	Value / Observation
Edge MCU Compute vs. GPUs	$\sim 2\times$ the total compute capacity of all GPUs worldwide.
Transformer First Token Delay	Tens to hundreds of milliseconds per token at small batch sizes—limited by DRAM.
Knowledge Retrieval Hit Rate	Up to 80% of typical search queries answerable on-device.
MCU DRAM Power Draw	Hundreds of milliwatts, compared to 5–10 mW for SRAM.
LLM Weight Footprint	Tens of MB, requiring $>$32 MB DRAM—unsuitable for MCUs with $<$1 MB SRAM.

TABLE V: Components, their primary functions, and typical use cases

Component	Primary Function	Use Case
Oryon CPU	Immediate processing needs	Real-time AI assistant responses
Adreno GPU	Image and video processing	Photo/video editing, live translation
Hexagon NPU	Sustained and pervasive AI applications	Generative AI, multi-modal tasks

- Edge Deployment: Facilitating the integration of LLMs into smart devices such as speakers, cars, and IoT systems.

The speech was also about the difficulties that engineers at Meta experience when deploying LLMs on embedded AI accelerators such as Qualcomm Hexagon and Apple Neural Engine (ANE):

- Memory Constraints: Devices often have a memory limit of approximately 4 GB, necessitating model partitioning.
- Dynamic Shapes: LLMs exhibit variable prompt and output lengths, which conflict with the static shape requirements of many accelerators.
- Quantization Limitations: Static quantization of activations and weights must be performed in advance, limiting flexibility.
- Operator Support Gaps: Certain tensor operations lack full kernel support, complicating implementation.
- Opaque Runtimes: Limited visibility into low-level scheduling and performance hinders optimization efforts.

In Table VI are shown all the principal optimization techniques developed to address these challenges.

Optimizing LLMs for hardware accelerators has profound implications for edge computing. By enabling on-device inference, these optimizations reduce reliance on cloud infrastructure, thereby lowering costs and improving privacy. Additionally, reasoning models, which require longer context windows and generate more tokens, can now be supported on edge devices. This advancement is crucial for applications such as autonomous agents, which rely on LLMs for decision-making and action orchestration.

Through techniques such as model partitioning, block-wise prompt encoding, KV cache management, and layer splitting, LLMs can be efficiently adapted to run on modern AI accelerators. These advancements pave the way for responsive, energy-efficient generative experiences on edge devices, reducing dependency on cloud resources and enabling a new era of intelligent, autonomous systems.

Bridging the Digital Divide Caused by Generative AI through the Edge

Generative AI, while revolutionary, has exacerbated the digital divide. Training LLMs with over 10^{14} parameters requires hundreds of millions of dollars and multi-gigawatt data centers, often powered by nuclear energy. This results in significant environmental costs, including 2.5 liters of water per kilowatt-hour of nuclear energy output and an estimated 2.5 million tons of e-waste by 2030. Only a handful of companies can afford these resources, leaving small and medium enterprises (SMEs), universities, and independent developers excluded from the GenAI ecosystem.

Edge AI offers a solution to this disparity by enabling generative tasks to run locally on devices without relying on expensive cloud GPUs. Frameworks like Tiny ML Foundation empower hobbyists, startups, and SMEs to participate in the generative AI revolution. This shift not only reduces costs but also minimizes environmental impact by eliminating the need for large-scale data centers. In Table VII are shown the results of a survey carried out by ST over 135 papers from 2022-2024 [Giorgetti and Pau(2025)] on the deployment of GenAI at the Edge, 66 of which were found to be in scope.

STMicroelectronics has demonstrated the functional feasibility of generative AI at the edge through hands-on demos:

- STM32N6: Integrated NPU (3 TOPS/W) achieves 10 FPS for neural style transfer on 320×320 images.
- STM32MP2: Dual-core ARM A15 with Linux supports:
  - LLM inference at 2–3 tokens/sec for a 0.5 B parameter model (Qwen).
  - A 1-bit quantized LLM (BitNet, last access 2025/7/10)) achieving a 4× speedup compared to int8 models.

TABLE VI: Optimization techniques, their descriptions, and benefits

Optimization Technique	Description	Benefits
Model Partitioning	Splitting deep LLMs (e.g., 32 layers) into sequential chunks to fit within accelerator RAM.	Enables deployment of large models on memory-constrained devices.
Block-wise Prompt Encoding	Processing prompts in blocks of tokens rather than one token at a time, leveraging KV cache.	Reduces latency and accelerates the time to generate the first token.
In-Place KV Cache Updates	Using strategies like shift pointers and smart masks to manage growing KV caches.	Maintains conversation history while minimizing memory overhead and maximizing throughput.
Layer Splitting for Apple NPU	Dividing large linear layers into smaller matrix multiplications and concatenations.	Improves compatibility with NPU limits and enhances processing speed.
Numerics Debugging Tooling	Comparing per-operator outputs on accelerators versus reference models to identify errors.	Speeds up debugging and ensures numerical accuracy in hardware environments.

TABLE VII: Summary of task categories, paper distribution, model sizes, and target devices

Task Category	% Papers	Model Size Range	Devices & Processors
Visual Processing	45%	0.4–1 B parameters	Smartphones (A15–A18, Exynos, Snapdragon); Jetson (15–50 W)
Image Generation	10%	0.4–1 B parameters	Smartphones (A17 Pro, Exynos NPU, A18 Pro, A16, Snapdragon 8 Gen 2)
Short Language Models	10%	125 M–4 B parameters	Smartphones
Text-to-Speech (TTS)	10%	Vocoder + TTS pipelines	Raspberry Pi, microcontrollers (RTF 0.01–2.0, real-time factor)
Vision QA & Other (¡5%)	35%	Varies (e.g., 428 M)	Jetson, Pi, microcontrollers (Tiny VQA on GAP8 @ low power; knowledge distillation used)

– A reasoning demo using the Moonshine Speech-to-Text (STT), Granite 3.3 and Piper for TTS connected into a pipeline, which transforms a thermostat into an intelligent edge agent. The pipeline stages are detailed in Table VIII:

Generative AI at the edge has the potential to bridge the digital divide by enabling broader participation in AI innovation. However, achieving this vision requires continued advancements in optimization techniques, hardware capabilities, and open-source collaboration. By addressing these challenges, edge AI can democratize access to generative AI, fostering a more inclusive and sustainable technological future.

Expanding Gen EdgeAI Horizon with Qualcomm Dragonwing AI on-prem Appliance Solution

The speech highlights the transformative potential of Qualcomm's DragonWing AI on-prem appliance solution in addressing the limitations of cloud-based GenAI while leveraging the strengths of edge computing. This discussion focuses on the technical aspects of the solution, emphasizing its architecture, use cases, and future directions.

Cloud-based GenAI solutions face significant challenges, including privacy concerns, high latency, unpredictable network dependencies, elevated energy consumption, and per-inference costs. Qualcomm's DragonWing AI appliance mitigates these issues by enabling large generative models (up to 70–100 billion parameters) to run locally with data center-class compute capabilities. This approach combines the benefits of edge computing and on-premises deployment, ensuring privacy, cost efficiency, and sustainability.

The DragonWing solution demonstrates versatility across multiple industries, enhancing productivity, safety, and customer engagement. Table IX summarizes the envisoned use cases and their associated benefits:

DragonWing is designed for high performance and compact deployment. Its architecture includes AI accelerator cards, a compact form factor, and a robust software stack:

- AI Accelerator Cards: Each card features 16 AI cores delivering 25 TOPS per core, achieving a peak performance of 400 FP16 TOPS. The cards support INT8, FP16, and FP32 precision, with future support for INT4 and INT1 for extreme quantization.
- Compact Form Factor: The appliance, with a footprint of approximately 20×20 cm, is fan-cooled, rack-mountable, and capable of running 70–100 billion parameter models like LLaMA 3 70B entirely on-premises.
- Software Stack: The Qualcomm AI Inference Platform supports over 1,000 open-source models and offers features like single-click deployment, autoscaling, and over-the-air updates. Developers benefit from SDKs, APIs, and a web-based playground for rapid prototyping.

DragonWing's performances are shown in Table X

Qualcomm's DragonWing solution is poised to evolve further, with next-generation SoCs targeting over 100 TOPS/W and deeper integration with sensor and robotic platforms. The scalable ecosystem, spanning IoT sensors to enterprise

TABLE VIII: Pipeline stages for devising a home thermostat agent with corresponding memory usage, latency, and notes

Pipeline Stage	Memory (MB)	Latency (s)	Notes
ASR (Moonshine)	65 MB	0.2× real time	Whisper alternative
LLM + Reasoning	750 MB	2–3 tokens/sec	Granite 3.3 model dominates runtime (98.5% execution time)
TTS (Piper)	185 MB	1× real time	Neural vocoder

TABLE IX: Industry applications with AI use cases and key benefits

Industry / Sector	AI Use Case	Key Benefits
Industrial / Oil & Gas	AI-assisted maintenance, predictive alerts	✓ Safety ↑ ✓ Productivity ↑ Downtime ↓
Retail / Hospitality	Conversational kiosks, personalized offers	✓ Customer engagement ↑ ✓ Revenue ↑
Logistics / Warehousing	Vision-powered defect detection, autonomous vehicles	✓ Accuracy ↑ ✓ Labor overhead ↓
Workforce Safety	Helmet/PPE compliance, hazard recognition	✓ Accident ↓ ✓ Compliance ↑
Training / Education	Interactive AR/VR labs, real-time tutoring	✓ Learning outcomes ↑ ✓ Instructors ↓

TABLE X: Key metrics and values for SoC performance and deployment

Metric	Value
SoC Performance	400 TOPS (16 × 25 TOPS cores), 500 MB on-chip SRAM per card
Model Scale	Up to 100 billion parameters at $<$1 second per token (FP16)
Latency & Throughput	5–10 tokens/sec for LLM inference (70B FP16); 1 second/image (512×512)
Energy Efficiency	$\sim$50 TOPS/W sustained (targeting $>$100 TOPS/W in next-gen SoCs)
Deployment	Over 50 pilot customers across manufacturing, retail, logistics, and life sciences

gateways, ensures broad applicability. Additionally, the focus on developer-friendly tools and open-source compatibility democratizes access to advanced AI capabilities.

The Qualcomm DragonWing AI on-prem appliance solution represents a significant leap in edge AI technology, addressing critical challenges in cloud-based GenAI while unlocking new possibilities across industries. Its compact design, robust performance, and developer-centric approach position it as a transformative tool for enterprises seeking privacy, cost efficiency, and sustainability in AI deployment.

AI Agents at the Edge - Architectures and Algorithms for Low-Latency Intelligence

The evolution of AI has reached a pivotal moment with the emergence of agentic AI systems operating at the edge. Unlike traditional deterministic edge AI models, modern agentic AI systems are capable of perceiving, generating, planning, remembering, and acting autonomously. This paradigm shift is driven by advancements in edge runtimes, agentic frameworks, and miniaturized models, enabling AI to function independently in low-connectivity environments while ensuring privacy and real-time responsiveness. This discussion explores the operational mechanisms of generative AI at the edge, its enabling technologies, and the potential for human-AI collaboration.

Traditional AI systems relied heavily on cloud-based infrastructures for static inference tasks such as classification and detection. However, modern edge AI agents operate directly on devices, offering significant advantages:

- Low/No Connectivity Scenarios: Edge AI is indispensable in remote clinics, disaster zones, and autonomous factories where network access is limited or unavailable.
- Privacy and Security: Data remains on the device, reducing exposure to external threats.
- Latency and Reliability: Real-time decision-making is achieved without network delays, ensuring consistent performance.

This shift underscores the importance of deploying AI systems that are not only efficient but also autonomous and secure.

The speech presented five core capabilities that distinguish Edge AI agents from traditional models:

- Perception: Local computer vision (CV) and ASR systems process data streams in real time. Examples include YOLO detectors for CV and Whisper CPP for ASR.
- Generation: Leveraging lightweight language models (e.g., Tiny Llama, Gemini 2B), agents produce natural language outputs.
- Planning: Multi-step workflows are executed using frameworks like Langraph and Autogen.
- Memory: On-device vector databases (e.g., Faiss, SQLite) enable RAG for context-aware responses.
- Tool Use: Agents interact with local APIs, scripts, or hardware to perform tasks such as triggering alerts or actuating motors.

These capabilities enable edge AI agents to act autonomously, adapting to dynamic environments and user needs.

The operational efficiency of edge AI agents is supported by advancements in model compression, runtime optimizations, and hardware capabilities. Key highlights are shown in Table XI:

Edge AI agents are transforming industries by enabling autonomous operations in challenging scenarios, some examples include:

- Disaster Response Drones: Equipped with on-device CV, drones analyze terrain changes and plan flight paths autonomously, achieving CV inference in $\leq$ 100 ms.
- Factory Inspection: Vision transformers detect defects, while language models generate reports and control robotic arms, operating with 1 billion parameter CV models.
- Personalized Retail Advisors: Agents detect empty shelves and generate real-time promotions using TTS systems, completing tasks in $<$ 200 ms.

These use cases highlight the versatility and impact of edge AI agents in real-world settings.

The deployment of edge AI agents necessitates robust security measures to protect sensitive againt data breaches in agent-to-agent (A2A) communication and ensure compliance. Key strategies include:

- Encrypted Local Stores: Securely store data on devices using hardware-based encryption (e.g., TPM, TrustZone).
- Tool Call Whitelisting: Restrict agent actions based on predefined conditions to prevent unauthorized operations.
- Audit Logging: Maintain accountability by logging all agent decisions and actions with timestamps.
- Hardware Root of Trust: Verify the integrity of models and runtimes through secure boot mechanisms.

The emergence of agentic AI at the edge represents a transformative shift in how AI systems operate. By leveraging advanced technologies, robust security protocols, and collaborative frameworks, edge AI agents are poised to revolutionize industries ranging from healthcare to manufacturing. As we continue to refine these systems, the focus must remain on enabling autonomy, ensuring security, and fostering meaningful human-AI collaboration.

NANDA: Towards building the internet of AI Agents

The speech by MIT Media Lab introduces NANDA, a transformative concept aimed at creating a decentralized "Web of AI Agents." This vision seeks to enable trillions of AI agents to interact autonomously, securely, and seamlessly, much like users navigate the World Wide Web (WWW) today. The discussion highlights the qualitative and quantitative aspects of NANDA, emphasizing its potential to redefine AI interactions and governance. The evolution of AI parallels the historical progression from mainframes to personal computers and eventually to the internet. Early AI models, such as ChatGPT and Gemini, function as standalone systems akin to PCs. However, the current trend involves interconnected agents operating in an intranet-like manner. NANDA aspires to establish an open "Web of Agents," where agents are autonomous, proactive, and capable of reasoning and taking actions. Unlike static web pages, these agents can perform tasks such as booking services, managing supply chains, or conducting augmented reality patrols.

The ultimate goal of NANDA is to create a decentralized ecosystem where trillions of edge and cloud-based AI agents can discover, authenticate, transact, and collaborate without centralized control. This vision introduces several challenges:

- Discovery: Dynamically locating unknown agents.
- Authentication and Trust: Verifying agent capabilities in a trustless environment.
- Transaction and Economics: Facilitating value exchange through "intelligence markets."
- Interoperability: Establishing seamless peer-to-peer communication protocols.
- Governance: Addressing privacy, user experience, and data rights.

The NANDA registry is designed to support the scalability and functionality of the agentic web. Its architecture is shown in Table XII

This architecture ensures scalability by maintaining a minimal registry size while decentralizing rich metadata. Post-discovery, agents communicate directly using protocols such as MCP or A2A.

TABLE XI: Technologies, details, and key metrics for edge AI deployment

Technology	Details	Metrics
Miniature Foundation Models	Models with 1–3 billion parameters (e.g., FEEL 1.3B) achieve GPT-3.5-level performance.	Quantization reduces memory usage by 4–8×, enabling deployment on 4–8 GB VRAM.
Runtimes & Toolchains	Tools like llama.cpp and WebGPU optimize local deployment.	FlashAttention achieves 10+ tokens/s on Jetson devices.
Hardware Examples	Devices like Dell laptops and Jetson Orin boards support edge AI workloads.	Up to 16 GB GPU VRAM enables efficient processing.

TABLE XII: Layers, roles, and key attributes of the NANDA Registry architecture

Layer	Role	Key Attributes
Lean Registry	Maps Agent ID to Metadata URL	Low-churn index for identifiability and discoverability.
Agent Facts	Decentralized metadata at each agent	Includes endpoints, skills, certifications, and security policies for trust/audit.
Dynamic Routing	Multi-endpoint resolution	Supports geo/load balancing, TTL, and versioning for global agent access.

The practical applications of NANDA were showcased through two key demonstrations:

- Agent-to-Agent Chat: A conversational exchange between two agents, Mary and Adam, where the tone of a daughter's request ("Can I watch a movie?") is rephrased to make it more agreeable.
- Agent-to-Entity Ordering: Adam's agent interacts with a bakery agent to negotiate and finalize a cake order. The process involves inventory checks and scheduling, all conducted autonomously over A2A protocols.

The NANDA initiative is set to advance further through development and collaboration. The open-source launch marks a significant milestone, as the NANDA registry code, metadata schema, and demo applications are now accessible to the public. Looking ahead, research efforts will delve into areas such as incentive design, reputation systems, private transactions, and user interface paradigms tailored for agentic applications. Additionally, the initiative extends an open invitation to researchers and industry professionals to join in shaping the democratic Web of AI Agents.

NANDA represents a significant leap toward a decentralized, agentic AI ecosystem. By addressing challenges such as discovery, trust, and interoperability, it lays the groundwork for a future where AI agents operate autonomously and collaboratively. The initiative's open-source approach and emphasis on scalability make it a promising framework for the next generation of AI interactions.

Continues integration with Executorch - the way LiquidAI utilizes PyTorch + Executorch for getting the optimal frontier

The speech delves into the intricacies of deploying machine learning models to edge environments using PyTorch and Executorch, and expands on the role of Executorch in achieving optimal performance.

The "edge" is philosophically defined as any environment devoid of a full machine learning runtime, such as PyTorch or CUDA. This environment is characterized by constrained compute and memory resources, offline operation, and strict packaging constraints. Unlike cloud-based systems, edge environments demand tailored solutions where models are optimized for inference and decision-making rather than training or fine-tuning. The edge is consumer-driven, meaning the success of a model is measured by its usability and performance in the end-user application.

- Compute and Memory Limitations: Unlike the virtually infinite resources of cloud GPUs, edge devices operate under tight resource budgets.
- Offline Operation: Models must function without persistent internet connectivity, often within closed or minimal operating systems.
- Application-Driven Deployment: Models are embedded as part of a host application, requiring external lifecycle management and no administrative privileges (e.g., no sudo commands).

Deploying models to edge devices involves addressing significant fragmentation across platforms. Each platform, such as Android ARMv8, iOS arm64, or Raspberry Pi aarch64, has unique CPU architectures, preferred runtimes, and operating system quirks. This results in a combinatorial explosion of builds, such a fragmented landscape necessitates the creation of dozens of distinct binaries, even before incorporating additional features.

The speaker also discusses the characteristics that a robust edge solution for GenAI should encompass:

1) Input Preprocessing and Tokenization: Custom and multimodal approaches may be required.
2) Grammar Constraints: Forcing specific outputs, such as JSON, by filtering logits.
3) Sampling Strategies: Advanced techniques beyond standard Top-k or Top-p sampling.
4) Custom APIs: Supporting diverse interfaces like CLI, REST, or gRPC, especially in environments where HTTP is unavailable.
5) Model Distribution: Ensuring models are packaged, encrypted, and ready for on-device deployment.
6) On-Device Fine-Tuning: RAG workflows for dynamic adaptation.
7) State Management: Handling checkpoint loading, key-value caching, and multi-pass conversations.

Liquid AI's innovative approach showcases remarkable efficiency improvements in various aspects of model development and deployment. In terms of build matrix capabilities, Liquid AI supports over 10 platform flavors. Each build process is completed in approximately 2 minutes, a stark contrast to the 9 hours and $200 cost associated with a single, private Lama CPP pull request. This efficiency highlights the platform's ability to streamline workflows and reduce operational expenses significantly.

Memory slicing is another area where Liquid AI demonstrates its technical prowess. For a 1 billion parameter model, the total weights amount to approximately 1 GB, while the mutable state requires only about 70 MB. Similarly, for a 3 billion parameter model, the total weights exceed 2 GB, and the mutable state requires around 200 MB. These optimizations ensure that even large-scale models can be managed effectively within constrained memory environments. The inference cost is also notably reduced. For instance, a Raspberry Pi 5 can run a 1 billion parameter model at a rate of approximately 10 tokens per second on a CPU using XNNPACK. This capability underscores Liquid AI's commitment to enabling high-performance inference on cost-effective hardware.

Executorch plays a pivotal role in bridging the gap between scientific model development and efficient edge deployment. It introduces efficient memory management techniques, such as caching mechanisms, which reduce runtimes from minutes to mere seconds. Additionally, Executorch supports custom operations tailored to specific hardware requirements, ensuring optimal performance across diverse platforms. The platform also provides a unified API, implemented as a lightweight C/C++/Rust layer. This API accelerates the loading of ONNX/FlatBuffer graphs compared to full framework exports. It integrates seamlessly into token loops, enabling advanced functionalities such as grammar-based logit filtering and custom sampling. Furthermore, the API exposes a clean C interface with higher-level bindings, facilitating build-once, run-anywhere functionality. This approach allows for the integration of custom operators, quantizers, and post-processors, as well as the orchestration of model loading and unloading, power modes, and thread priorities.

Liquid AI's comprehensive solutions exemplify its dedication to advancing the state of AI technology while ensuring accessibility and efficiency for a wide range of users and applications.

Deploying machine learning models to edge environments is a complex process requiring meticulous optimization and hardware-specific tailoring. Liquid AI's Executorch simplifies this process by providing a unified, model-agnostic core that compiles efficiently into minimal C APIs for diverse targets. This approach allows teams to focus on enhancing model quality and application integration rather than grappling with cross-compilation and packaging challenges.

Language Models at the Edge, From Feasibility to Collaboration

The push to deploy GenAI at the edge is driven by a combination of technical, economic, and user-centric factors. Centralized LLMs face persistent challenges, including high provisioning costs, security and privacy risks, and vulnerability to service disruptions. These limitations have catalyzed a shift toward smaller, smarter, and closer AI systems, referred to as SLMs. This discussion explores the feasibility, trade-offs, and collaborative potential of edge-based AI deployments.

The edge paradigm offers compelling advantages in privacy, latency, and cost-effectiveness. Centralized LLMs retrain on user queries, heightening the risk of exposing sensitive data. Additionally, immersive and video-powered AI assistants amplify latency issues when relying on centralized services. The emergence of smaller, open-source models with fewer than 10 billion parameters has further enabled the feasibility of edge AI. These trends collectively underscore the need for a "smaller, smarter, closer" approach to AI deployment.

While smaller models are more efficient, they often exhibit higher rates of hallucination. For instance, a 0.5 billion parameter model may generate inaccurate or fabricated responses, whereas an 8 billion parameter model demonstrates significantly higher accuracy. Specialization through fine-tuning can mitigate this issue. For example, fine-tuning a 1.5 billion parameter model on a narrow dataset can achieve accuracy levels comparable to a 72 billion parameter cloud model, demonstrating the potential of SLMs for specific applications. Table XIII summarizes the performance metrics for various edge devices and models.

As shown by Table XIII:

- GPUs consistently deliver lower latency compared to CPUs, with a 2–5× speed advantage on devices like the Jetson AGX Orin.
- Quantization (e.g., 8-bit to 4-bit) reduces inference time and energy consumption, with up to 70% energy savings in CPU mode.

Edge deployments often achieve higher performance-cost ratios (PCRs) compared to cloud-based solutions. The PCR metric combines response quality, responsiveness, and operating cost. While edge systems may exhibit slightly lower accuracy, their near-zero API fees and local execution make them cost-effective. For instance, the PCR for edge AI im-

TABLE XIII: Device performance metrics including model size, inference latency, and energy savings

Device	Model Size (B params)	Inference Latency (GPU vs. CPU)	Energy Savings ($8 \rightarrow 4$-bit Quantization)
Raspberry Pi 5	0.5–9.2	GPUs 2–5× faster	CPU: $10\% \rightarrow 70\%$; GPU: $20\% \rightarrow 40\%$
Jetson Nano	0.5–9.2	GPUs 2–5× faster	CPU: $10\% \rightarrow 70\%$; GPU: $20\% \rightarrow 40\%$
Jetson AGX Orin	0.5–9.2	GPUs 2–5× faster	CPU: $10\% \rightarrow 70\%$; GPU: $20\% \rightarrow 40\%$

plementations is significantly higher when energy costs are factored in, as shown in the equation:

$$PCR = \frac{U}{CPR} \tag{1}$$

where

- $U = \alpha Q + (1 - \alpha)R$ (the utility function U combining quality Q and responsiveness R)
- $CPR_{cloud} = API_{cost}$
- $CPR_{edge} = \text{Energy}(kWh) \times \text{Electricity Rate}(c/kWh)$

Edge clusters comprising devices like the Raspberry Pi 5, Jetson Nano, and Jetson AGX Orin can handle varied query loads effectively. Task assignment policies, such as load-aware allocation, minimize cloud fallback and optimize device utilization. For example, dynamic load-aware strategies outperform random or capacity-prioritized approaches in balancing workloads during bursty query patterns.

The speech also highlights how the future of edge AI lies in hybrid edge–cloud workflows and agentic collaboration. In this paradigm, edge nodes and cloud services coordinate to:

- Offload computation selectively.
- Share knowledge across agents.
- Decompose complex queries into subtasks.

This multi-agent fabric enables a balance between local inference speed, cloud-scale quality, and robust privacy controls.

Deploying generative LLMs on edge devices is increasingly feasible, but it requires navigating trade-offs in model size, quantization, and energy efficiency. By leveraging unified performance metrics and orchestrating tasks across edge clusters, organizations can achieve cost-effective and responsive AI solutions. The integration of agentic collaboration further enhances the potential of edge AI, paving the way for scalable and privacy-conscious deployments.

Agentic AI with SLMs for Airport Boarding Process Monitoring

The implementation of agentic AI systems for real-world problems, such as monitoring airport boarding processes, represents a significant advancement in operational efficiency. This presentation highlights the development of an edge-based AI system designed to address delays caused by passenger behavior during boarding. This discussion explores the qualitative and quantitative aspects of the system, its architecture, and its potential impact on airport operations.

The boarding process at airports often faces inefficiencies due to passengers not adhering to designated zones, mixing with other flights, or wandering off. These behaviors can extend boarding times for up to 150 passengers to 30–60 minutes, leading to flight delays. The stakeholders involved in addressing this issue include Turin Airport operations, airport handlers, low-cost carriers, STMicroelectronics, and Eurix. The goal is to deploy an automated system that monitors boarding bottlenecks in real time and alerts staff to reduce delays.

The agentic AI system developed by Andrea's team is designed following a multi-agent, multi-layer architecture to enable real-time monitoring and decision-making. This architecture integrates several key components, each contributing to the system's overall functionality. These agents operate on Clearpath Jackal robots, which are equipped with RTX 4070 GPUs and STM32MP2 microcontrollers to facilitate edge deployment. The CV capabilities of the system are powered by a YOLO-based model, optimized for person detection, crowd grouping, and counting at a rate of 30 frames per second. To enhance accuracy, a custom formation algorithm is employed to filter transient clusters, thereby reducing the likelihood of false alarms.

Furthermore an aggregate programming layer plays a critical role in ensuring fault tolerance and achieving consensus among the agents. This layer addresses potential operational challenges, such as low battery levels or camera occlusion, to maintain system reliability. The agentic LLM interpreter is another essential component of the architecture. Initially tested on laptops equipped with RTX 4070 GPUs, the system is planned to be migrated to Jetson Nano and STM32 platforms for edge deployment. The interpreter processes outputs from the CV system and associated metadata to classify situations as either critical or non-critical, typically within a processing time of 10 to 20 seconds.

Finally, the consensus and alerting mechanism ensures timely communication of critical information. When multiple agents detect a critical state, the system dispatches alerts to staff through a local network user interface. This mechanism ensures that critical situations are promptly addressed. The system's performance is characterized by the metrics presented in Table XIV

The development of the agentic AI system required addressing several engineering challenges. These challenges included integrating heterogeneous frameworks, optimizing resource constraints, refining model parameters, and managing thermal and battery performance.

- The integration of heterogeneous frameworks involved

TABLE XIV: Key metrics and values for real-time CV robot deployment

Metric	Value
Real-time CV performance	30 FPS on edge GPUs; 15–20 FPS on microcontroller prototypes
Detection radius per robot	$\leq$ 20 m
Robots per gate area	3–5
Detection-to-decision latency	$<$ 2 seconds (CV: $<$ 50 ms, aggregate check: $<$ 100 ms, LLM inference: $\sim$1–2 s)
Deployment runtime	Continuous 1-hour operation on battery
Deployment target	Live demo on July 15 at Gates 14–17, handling up to 150 passengers

combining ROS 2, custom C/C++ computer vision code, and llama.cpp for LLMs. This was achieved using lightweight local remote procedure calls (RPC), ensuring seamless communication between components.

- Resource constraints posed significant challenges, particularly in optimizing the system for power-efficient ARM cores and managing the heat and power profiles of the Jetson Nano. Additionally, the system had to operate within the memory limitations of 1 GB on microcontrollers.
- Model optimizations were critical to achieving efficiency. Aggressive pruning and 4-bit quantization reduced the LLM parameters from 2 billion to 0.5 billion. Sparse attention mechanisms and the use of domain-specific vocabulary further enhanced the model's performance.
- Thermal and battery management were also key considerations. The system was designed to ensure minimal onboard cooling requirements while maintaining a patrol time of at least one hour without recharging.

In conclusion, the agentic AI system for airport boarding monitoring exemplifies the integration of advanced AI techniques with real-world applications. By leveraging edge computing, optimized SLMs, and distributed decision-making, the system is able to address critical operational challenges, offering a scalable and efficient solution for crowd-controlling in modern airports.

A Survey by Generative EDGE AI Workin Group

As part of its commitment to inclusive innovation and global collaboration, the Generative EDGE AI WG has launched a strategic community survey. Initially shared among the 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.

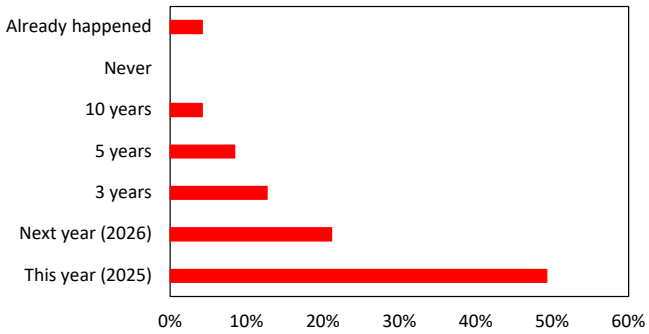

Fig. 2: Market outlook survey: When do you expect Edge Generative AI solutions to start reaching the market?

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

Survey Highlights

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

Market timing (Fig. 2) expectations are optimistic: a clear majority of respondents (over 70%) anticipate that GenEdgeAI 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 (Fig. 3), the community showed strong interest in SLMs, Visual Question Answering (VQA), STT, and TTS 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 (Fig. 4), respondents expect to see impact across the stack: hardware/chips, applications, and

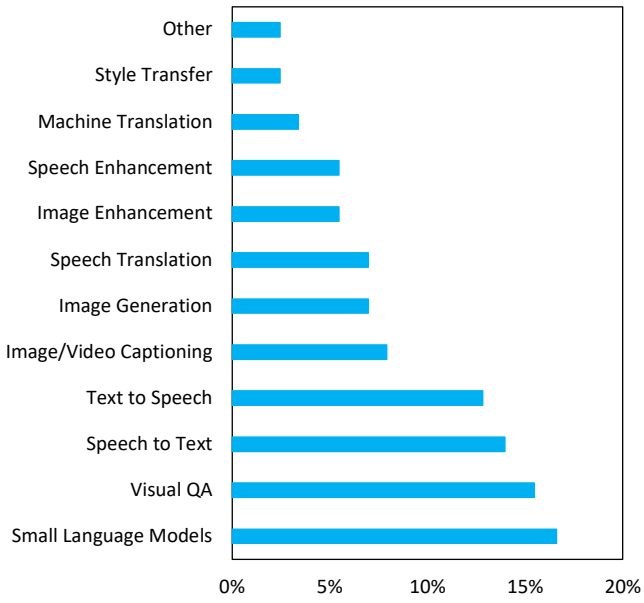

Fig. 3: Application priorities survey: Which use cases for Edge Generative AI are most relevant or promising?

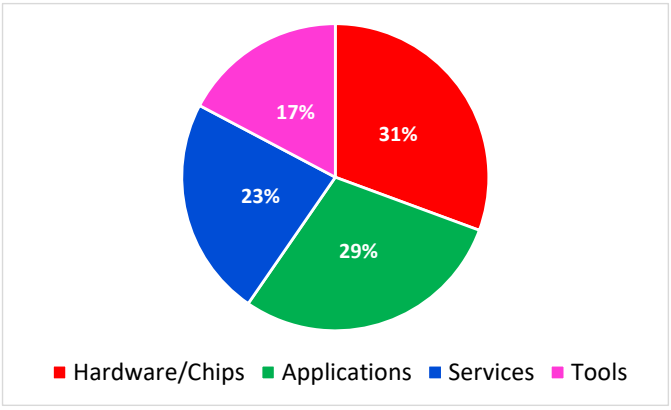

Fig. 4: Solution types: Which types of Edge Generative AI solutions are most likely to emerge?

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 GenEdgeAI.

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 research and development (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

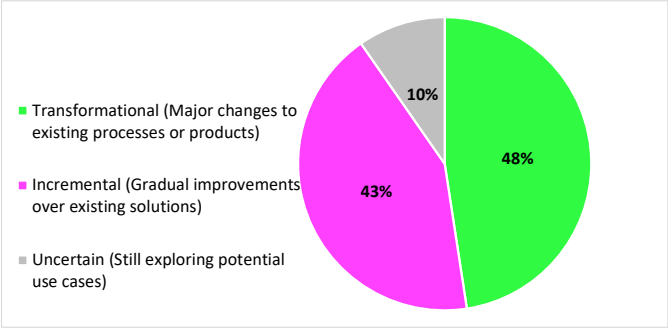

Fig. 5: Industry impact: How significant do you expect the impact of Edge Generative AI to be in your industry?

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 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 GenAI is overwhelmingly positive (Fig. 5), 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 (Fig. 6), 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.

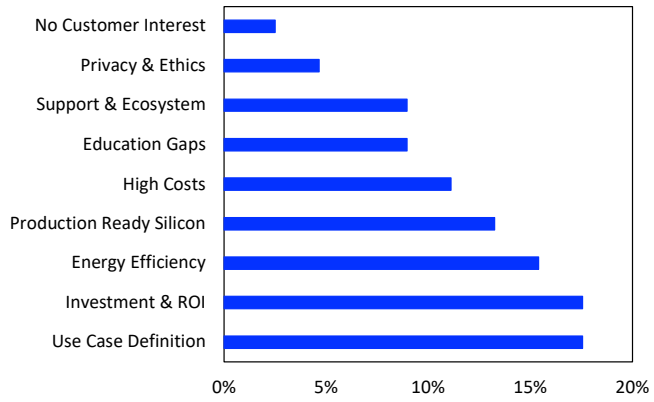

Fig. 6: Adoption challenges: What are the main barriers to adopting Edge Generative AI in your organization?

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

The WG is also considering opening the survey to the broader public, already active in GenEdgeAI or just beginning to explore its potential. These inputs can help steer the group's direction and ensure that its work is aligned with real-world challenges and opportunities.

Any voice matters, and together, the WG can build a stronger, more connected, and impactful GenEdgeAI ecosystem.

On Premises GenAI

There is an increasing demand from small and medium businesses, enterprises and industrial organizations to run custom and off-the-shelf GenAI applications at their premises. Running AI learning and inference on premises can deliver significant savings in operational costs compared to the cost of renting Cloud AI infrastructure. Using the On-Prem GenAI, the community can now ensure data ownership and control the technology infrastructure to build and offer its own models customized services.

Get Involved

Whether you're developing models, building systems, optimizing hardware, or exploring novel applications, the GenEdgeAI WG 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.

Acknowledgment

Conceptualization, D. Pau, H. Moussanif; methodology, D. Pau, H. Moussanif; validation, D. Pau, H. Moussanif; formal analysis, R. Morabito, D. Pau, H. Moussanif; investigation, D. Pau, H. Moussanif; resources, D. Pau; data curation, D. Pau, H. Moussanif; writing—original draft preparation, R. Morabito, R.Adorante, D. Pau; writing—review and editing, R.Adorante, D. Pau; visualization, R. Morabito, H. Moussanif, R.Adorante, D. Pau; supervision, D. Pau; project administration D. Pau; funding acquisition, D. Pau. All authors have read and agreed to the published version of the manuscript.

II. Abbreviations

The following abbreviations are used in this manuscript:

ADAS	Advanced Driver-Assistance Systems
A2A	Agent to Agent
AI	Artificial Intelligence
ASR	Automatic Speech Recognition
CPU	Central Processing Unit
DOAJ	Directory of open access journals
DSP	Digital Signal Processing
FAQ	Frequently Asked Questions
FP16	Floating Point 16 bits
GenAI	Generative AI
GenEdgeAI	Generative Edge AI
GPU	Graphics Processing Unit
IoT	Internet of Things
LD	Linear dichroism
LLM	Large Language Model
MAI	Multimodal Artificial Intelligence
MCU	Micro-Controller Unit
MDPI	Multidisciplinary Digital Publishing Institute
NPU	Neural Processing Unit
OPCW	Organization for the Prohibition of Chemical Weapons
PC	Personal Computing
PCR	Performance-Cost Ratios
PTQ	Post Training Quantization
RAM	Random Access Memory
RAG	Retrieval Augmented Generation
ROI	Return on Investment
RPC	Remote Procedure Calls
R&D	Research and Development
SLM	Small Language Model
SME	Small and Medium Enterprises
SoC	System-on-chip
STT	Speech-to-Text
TLA	Three letter acronym
TTS	Text-to-Speech
VLM	Visual Language Model
VQA	Visual Question Answering
WG	Working Group
WWW	World Wide Web
XR	Extended reality

References

[Morabito et al.(2025)Morabito, Pau, Moussanif] Morabito, R.; Pau, D.P.; Moussanif, H. Generative Edge AI Working Group: An Initiative of The Edge AI Foundation

[Soldatos and Rao(2024)] Soldatos, J.; Rao, R. Exploring the dynamic world of Edge AI applications across industries. In: Jaber, S. (Ed.), 2024, p. 55. Available online: https://dateurope.com/wp-content/uploads/2024/05/2024STAGEOFEDGEAIREPORT.pdf (accessed on July 2025).

[Giorgetti and Pau(2025)] Giorgetti, G.; Pau, D.P. Transitioning from TinyML to Edge GenAI: A Review. *Big Data and Cognitive Computing* **2025**, *9*(3), 61. https://www.mdpi.com/2504-2289/9/3/61. doi:10.3390/bdcc9030061.