# From Feasibility to Ecosystems: How Generative AI at the Edge Has Evolved

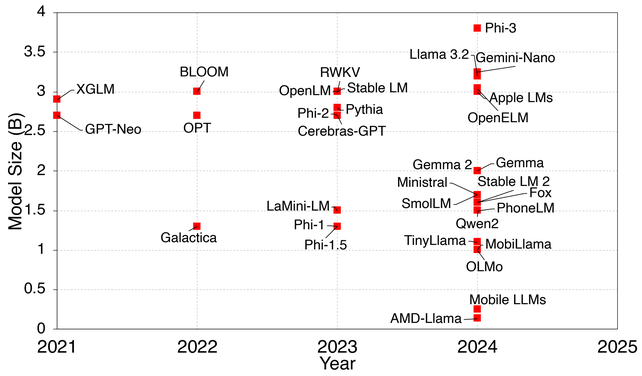
Roberto Morabito* (EURECOM) and Danilo Pau* (STMicroelectronics)

*Generative Edge AI Working Group Chairs

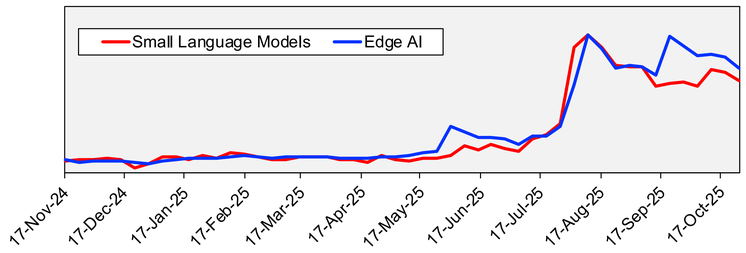
**Generative Edge AI is no longer about "can we run Small Language Models?"**
**It is about engineering reproducible and interoperable AI ecosystems at scale.**

## *Why Generative Edge AI*

Recent advances in ***Small Language Models*** (SLMs) have enabled execution on embedded and edge platforms.



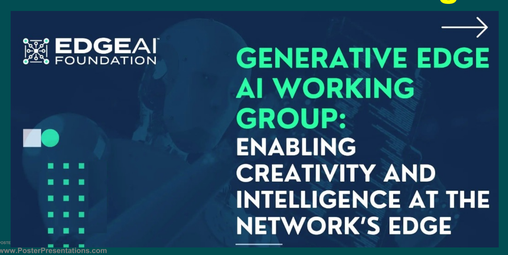

As SLM efficiency improved, interest in Edge AI and compact language models began to evolve in parallel.




**Feasibility became realistic.**

**System integration became the dominant bottleneck.**

**Read about the Working Group mission**




## *The Five Evolutionary Waves (2024–2025)*


1st Forum March 2024


2nd Forum October 2024


3rd Forum May 2025


4th Forum November 2025


5th Forum April 2026

**Watch the last Gen Edge AI forum!**

**2024**
**2025**

1. **FEASIBILITY: On-device** — "Can we run SLMs?"
2. **OPTIMIZATION: Edge-optimized pipelines** — compilers, CPU optimizations, toolchains.
3. **COORDINATION: Distributed edge intelligence** — multi-device inference, agentic systems.
4. **DOMAIN DEPLOYMENT: Verticalization** — Domain-specific deployments (e.g., automotive, healthcare).
5. **ECOSYSTEM ENGINEERING: Multimodal + Agentic Edge AI** — multimodality, orchestration, workflows.

***Generative Edge AI is transitioning from experimentation to ecosystem building.***

**Scan for reading our Papers!**

Paper 1

Paper 2

## *From Models to Ecosystems*

***Industry Signals***

*Survey insights from Edge AI Foundation partners*

- **Adoption timeline**: Near-term
- **Primary barriers**: Tooling & system integration
- **High-impact sectors**: Healthcare, Industrial

**Models are perceived as sufficiently capable.**

**Toolchains, orchestration and system integration are the limiting factors.**

***The Paradigm Shift***

| Model-Centric Era | Ecosystem-Centric Era |
|---|---|
| • Quantization & compression<br>• Latency benchmarks<br>• Single-device demonstrations<br>• Architecture comparisons | • Agent coordination across devices<br>• Interoperability across hardware<br>• Multimodal integration<br>• Reproducible deployment workflows |

**Progress is increasingly determined by the surrounding ecosystem – not raw model size.**

## *Implications and Research Directions*

- **Agentic Edge Systems**

Design **lightweight coordination protocols** that account for: *(i) Intermittent connectivity, (ii) Energy and memory constraints, (iii) Hardware heterogeneity*

- **Interoperability**

**Model and agent interaction** that is: *(i) Semantically aligned, (ii) Resource-aware, (iii) Hardware-conscious*

- **Multimodal and Physical Edge Intelligence**

**Integrate language** with *vision, audio, biosignals, and sensor streams, enabling Vision-Language-Action (VLA) models that connect perception, reasoning, and actuation in edge environments*

- **Reproducible Toolchains**

Bridge **research prototypes and production systems** through: *(i) Developer-friendly pipelines, (ii) Hardware abstraction, (iii) Deployment reproducibility*

**The next phase of Generative Edge AI will be defined by ecosystem engineering.**