Benchmarking Small Language Models on an Industry-grade, High-end Microcontroller

EDGEAI FOUNDATION Research Symposium 2026

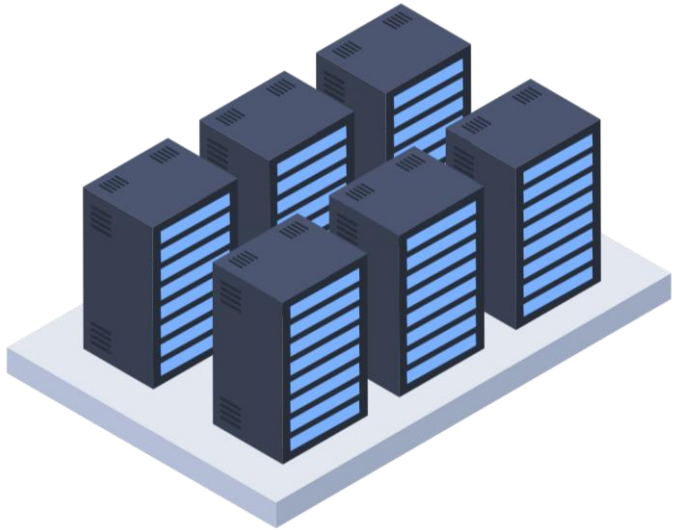
Pietro Firpo[1], Riccardo Berta[1], Francesco Bellotti[1], Luca Lazzaroni[1], Danilo Pietro Pau[2]

[1] University of Genoa

[2] STMicroelectronics

Introduction

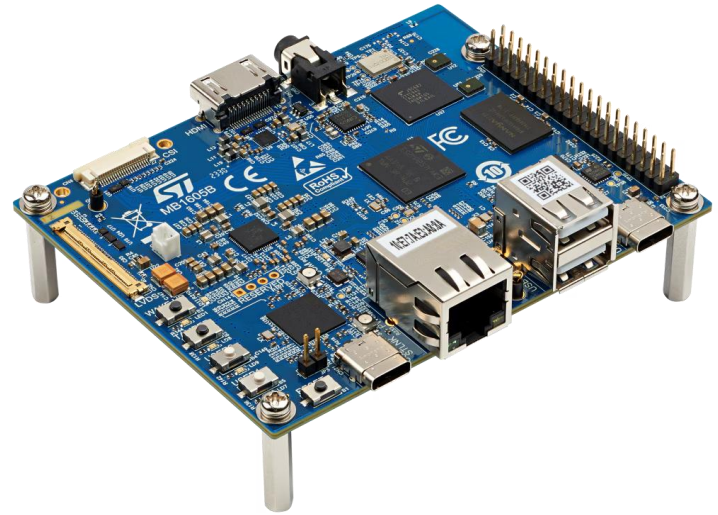
From the Cloud to the Edge



From LLMs to SLMs

- Size and HW requirements
- Benchmarking: accuracy-oriented vs latency oriented

Chosen a platform:



- **RQ1**: What inference throughput can SLMs achieve on an industry-grade MCU-MPU hybrid?
- **RQ2**: How does throughput scale with model size, and how predictive is parameter count of deployability?
- **RQ3**: Which throughput regimes correspond to viable embedded language-based applications?

State of the Art

SLMs:

- DistilBERT, ALBERT: smaller size, similar accuracy
- SLMquant, TensorSLM: compression techniques for LMs

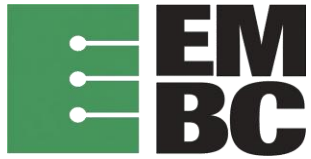
LM Benchmarking:

- GLUE: accuracy
- MLPerf: latency and throughput. High-end HW
- SLM-bench: throughput on Nvidia Jetson AGX Orin

MCU benchmarking:

- MLPerf Tiny: TinyML on MCUs
- EEMBC: embedded AI for signal processing or vision

Filling the gap: no data on LM latency/throughput on MCUs/MPUs



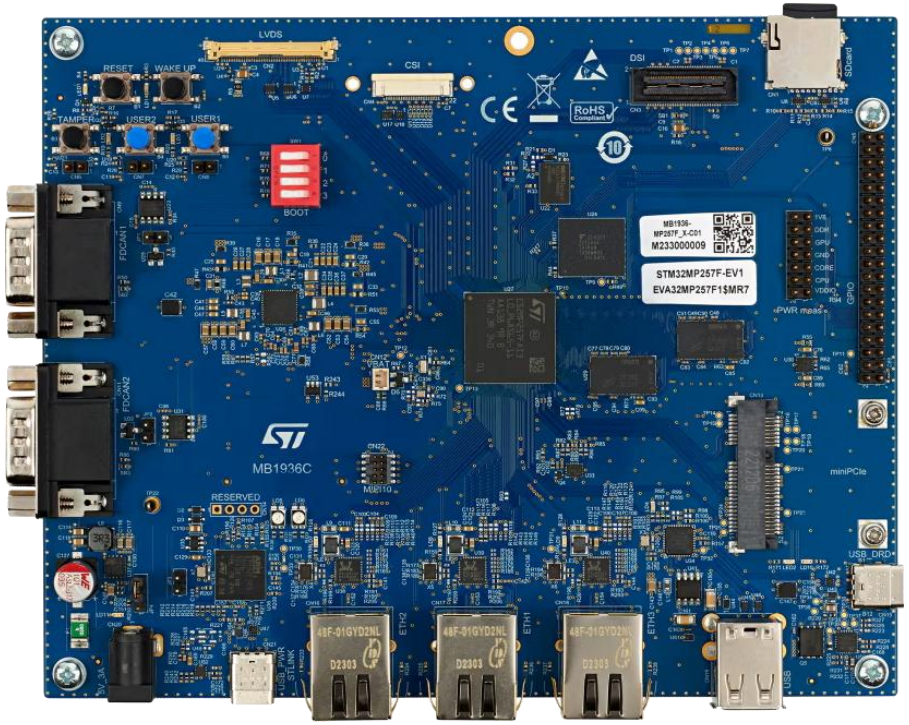
Experimental Settings

Metric: Inference Throughput

Platform: STM32MP257F-EV1

- Compromise between ultra low power and GPU powered

Component	Description
Application cores	2 × Arm Cortex-A35
Real-time core	Arm Cortex-M33
AI accelerator	VeriSilicon GC8000UL NPU
RAM	4 GB
Clock frequency	1.5 GHz



Inference frameworks:

- Llama2.c: weights and activations in FP32
- Llama.cpp: Q8 weights, Q16 activations
- Ollama: Q4 weights, Q8 activations

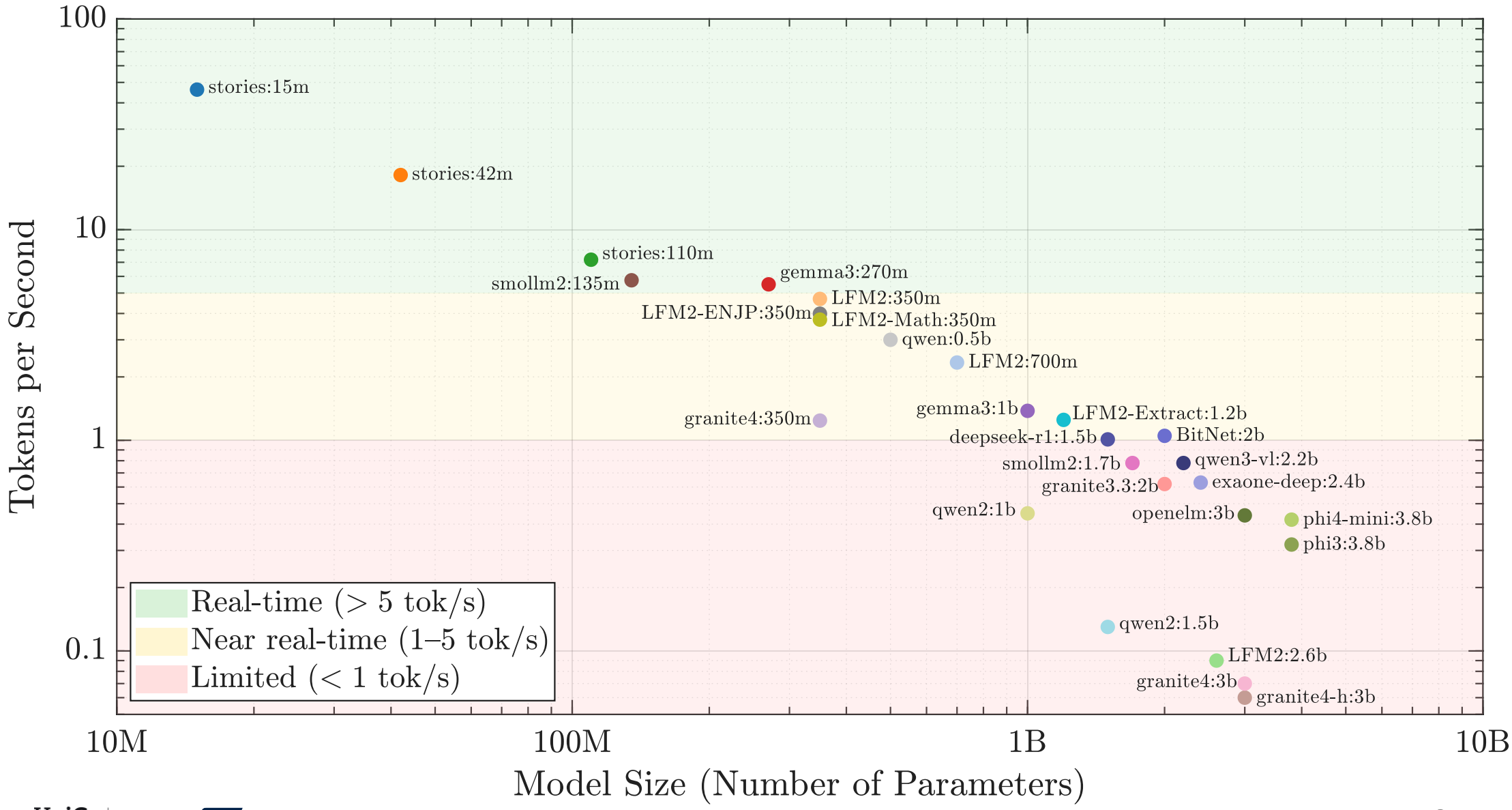
Experimental Settings

Models tested

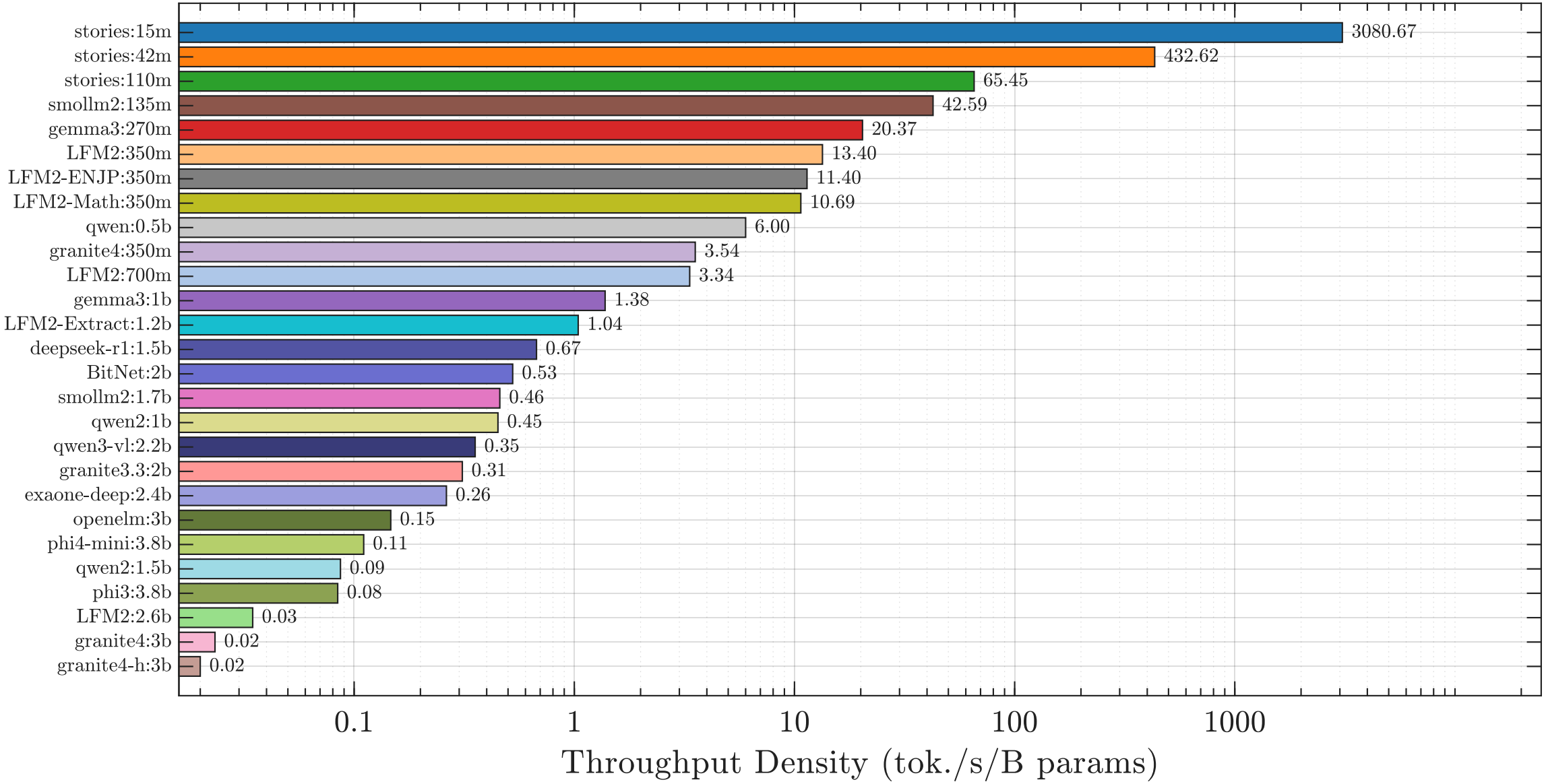
Model Name	General purpose	Params	Encoder layers	Inference framework
stories:15m	No (story generation)	15M	6	llama2.c
stories:42m	No (story generation)	42M	8	llama2.c
stories:110m	No (story generation)	110M	12	llama2.c
smollm2:135m	Yes	135M	30	ollama
smollm2:1.7b	Yes	1.7B	24	ollama
LFM2-ENJP:350m	No (EN–JP translation)	350M	16	llama.cpp
LFM2-Math:350m	No (math reasoning)	350M	16	llama.cpp
LFM2-Extract:1.2b	No (information extraction)	1.2B	16	llama.cpp
LFM2:350m	Yes	350M	16	ollama
LFM2:700m	Yes	700M	16	ollama
LFM2:2.6b	Yes	2.6B	30	ollama
deepseek-r1:1.5b	Yes	1.5B	28	ollama

Model Name	General purpose	Params	Encoder layers	Inference framework
gemma3:270m	Yes	270M	18	ollama
gemma3:1b	Yes	1B	26	ollama
granite3.3:2b	Yes	2B	40	ollama
granite4:350m	Yes	350M	28	ollama
granite4:3b	Yes	3B	40	ollama
granite4-h:3b	Yes	3B	40	ollama
qwen:0.5b	Yes	500M	24	ollama
qwen2:1b	Yes	1B	16	llama.cpp
qwen2:1.5b	Yes	1.5B	28	llama.cpp
qwen3-vl:2.2b	Yes	2.2B	28	ollama
BitNet:2b	Yes	2B	30	proprietary
exaone-deep:2.4b	Yes	2.4B	30	ollama
openelm:3b	Yes	3B	36	ollama
phi3:3.8b	Yes	3.8B	32	ollama
phi4-mini:3.8b	Yes	3.8B	32	ollama

Results


100
10
1
0.1
Tokens per Second
stories:15m
stories:42m
stories:110m
smollm2:135m
gemma3:270m
LFM2:350m
LFM2-ENJP:350m
LFM2-Math:350m
qwen:0.5b
LFM2:700m
granite4:350m
gemma3:1b
LFM2-Extract:1.2b
deepseek-r1:1.5b
BitNet:2b
smollm2:1.7b
qwen3-vl:2.2b
granite3.3:2b
exaone-deep:2.4b
qwen2:1b
openelm:3b
phi4-mini:3.8b
phi3:3.8b
qwen2:1.5b
LFM2:2.6b
granite4:3b
granite4-h:3b
Real-time (> 5 tok/s)
Near real-time (1–5 tok/s)
Limited (< 1 tok/s)
10M
100M
1B
10B
Model Size (Number of Parameters)

Results


stories:15m 3080.67
stories:42m 432.62
stories:110m 65.45
smollm2:135m 42.59
gemma3:270m 20.37
LFM2:350m 13.40
LFM2-ENJP:350m 11.40
LFM2-Math:350m 10.69
qwen:0.5b 6.00
granite4:350m 3.54
LFM2:700m 3.34
gemma3:1b 1.38
LFM2-Extract:1.2b 1.04
deepseek-r1:1.5b 0.67
BitNet:2b 0.53
smollm2:1.7b 0.46
qwen2:1b 0.45
qwen3-vl:2.2b 0.35
granite3.3:2b 0.31
exaone-deep:2.4b 0.26
openelm:3b 0.15
phi4-mini:3.8b 0.11
qwen2:1.5b 0.09
phi3:3.8b 0.08
LFM2:2.6b 0.03
granite4:3b 0.02
granite4-h:3b 0.02
0.1
1
10
100
1000
Throughput Density (tok./s/B params)

Conclusions

RQ1: inference throughput on hybrid MCU-MPU

Broad range, 0.1 tok./s < throughput < 10 tok./s

RQ2: throughput scaling wrt model size and relationship between param. count and deployability

Throughput influenced by more than size. Difficult scaling

RQ3: throughput regimes for applications

> 5 tok./s: RT and latency-sensitive applications

1-5 tok./s: near-RT and agentic

< 1 tok./s: severely limited, highly restrictive use cases

Università di Genova | SCUOLA POLITECNICA

EDGEAI FOUNDATION

Thanks for your attention