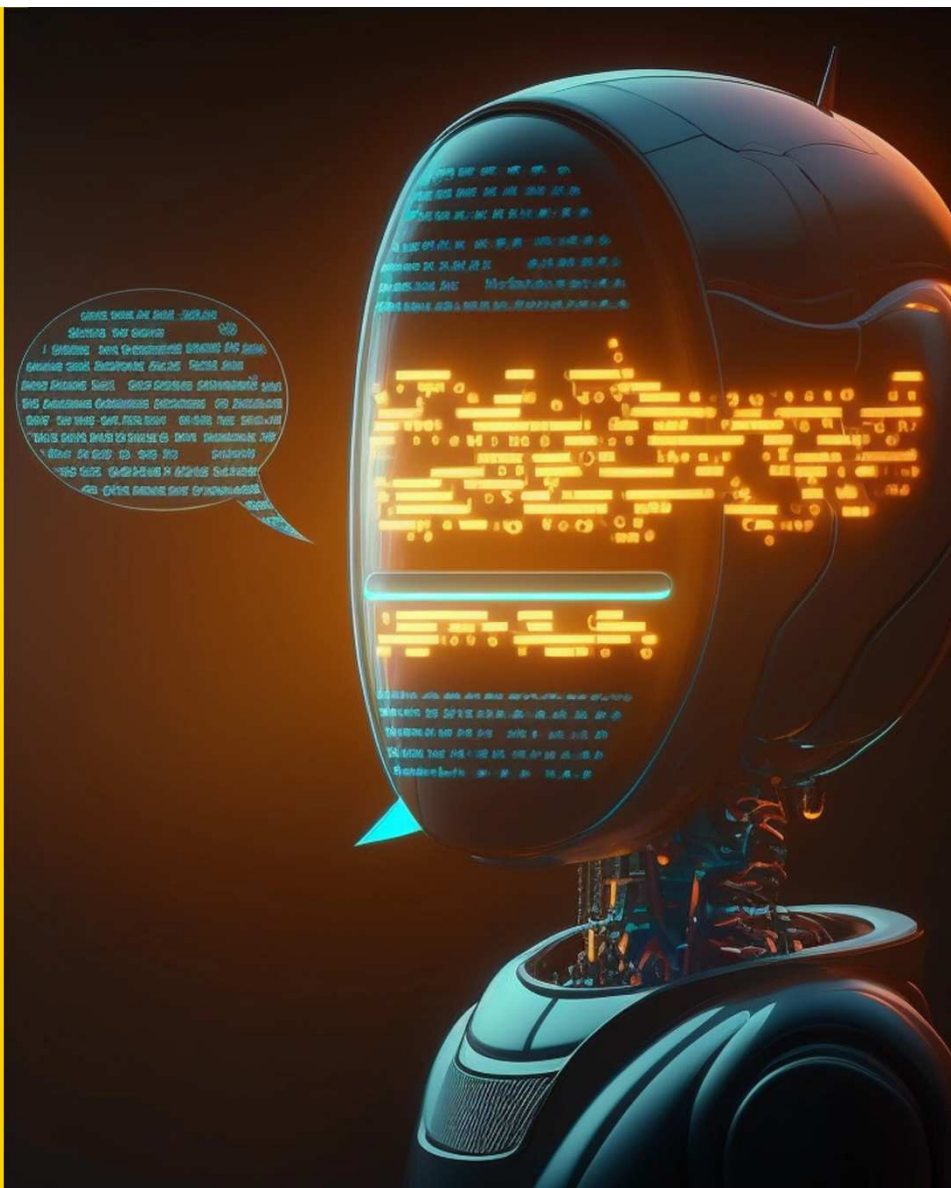


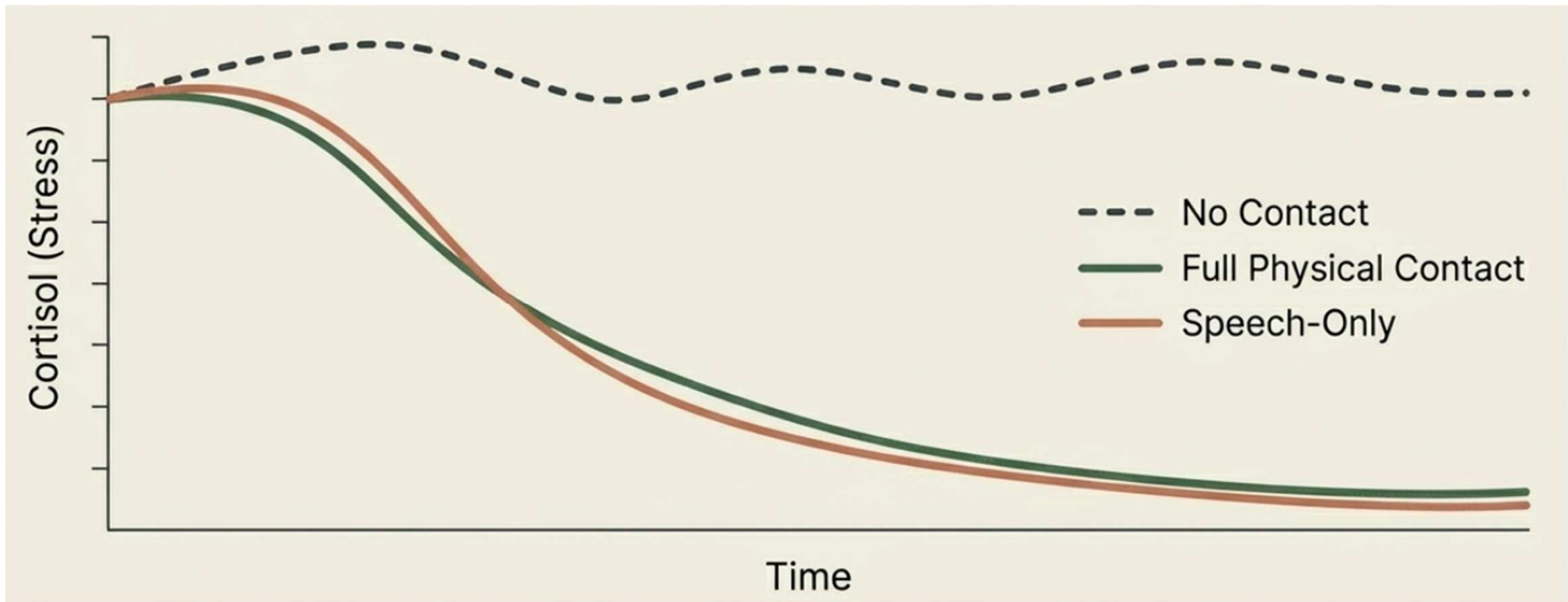


Conversational agents for natural human machine interaction: which type of edge device?

Technical Director, System Research
IEEE, AAIA, ST Fellow
Full professor ASN09/E3 Italy
IEEE Distinguished Industry Lecturer
Scientific Committee 2026-2028 IEEE SPS Italy
Global Governance and AI Safety NAAI Committee
APSIPA Life Member



Conversation is a touch at distance



Seltzer LJ, Prosofski AR, Ziegler TE, Pollak SD.
Instant messages vs. speech: hormones and why we still need to hear each other. *Evol Hum Behav.* 2012 Jan;33(1):42-45. doi: 10.1016/j.evolhumbehav.2011.05.004. PMID: 22337755; PMCID: PMC3277914.



Conversations for good humor

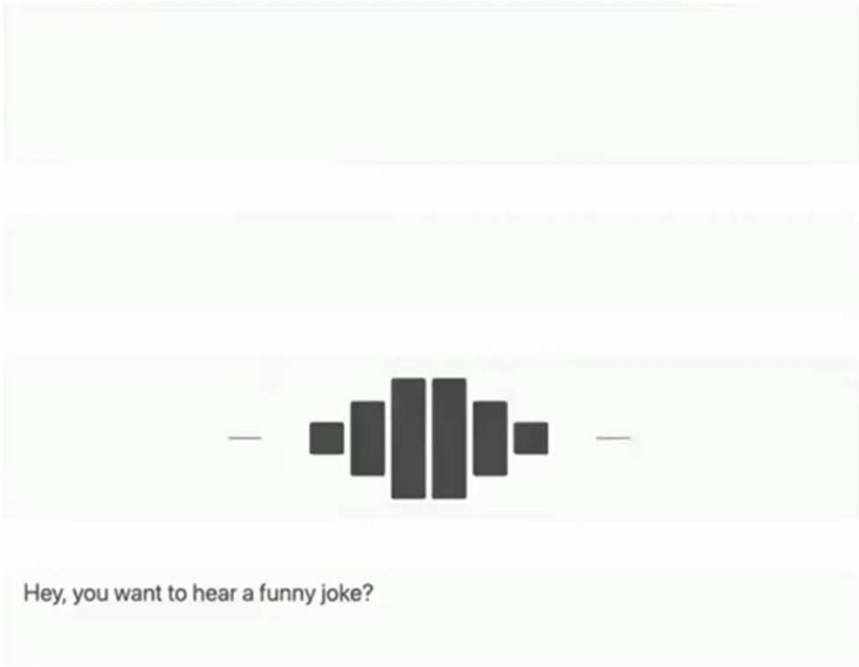


ASR→LLM→TTS

Built on **Moshi (7B)**
with Mimi
encoder/decoder
and **Helium LLM**;

NVIDIA
PersonaPlex

User



Conversational AI Gen EdgeAI

Perceptual AI (a.k.a tinyML)

Audio classification, ASR, Speaker ID
AED, Anomaly Detection, Audio separation,
Noise suppression
Speech intents understanding, Audio tagging,
Music retrieval

Scraping HF

- **audio**: ["audio gguf", "speech gguf", "sound gguf", "music gguf", "speech-to-text gguf", "speech to text gguf", "stt gguf", "asr gguf", "transcribe gguf", "transcription gguf", "audio classification gguf", "sound classification gguf", "audio tagging gguf", "sound event detection gguf", "keyword spotting gguf", "speaker recognition gguf", "speaker verification gguf", "speaker diarization gguf", "speaker embedding gguf", "audio separation gguf", "source separation gguf", "audio enhancement gguf", "speech enhancement gguf", "noise suppression gguf", "denoising gguf", "vocoder gguf", "audio codec gguf", # **TTS**
keywords "tts gguf", "text-to-speech gguf", "text to speech gguf", "speech synthesis gguf", "voice synthesis gguf", "voice assistant gguf", "text-to-audio gguf", "text to audio gguf", "audio generation gguf", "music generation gguf", **many more search keys**



Conversational AI Gen EdgeAI

Audio
493 models

Conversational
16,179 models

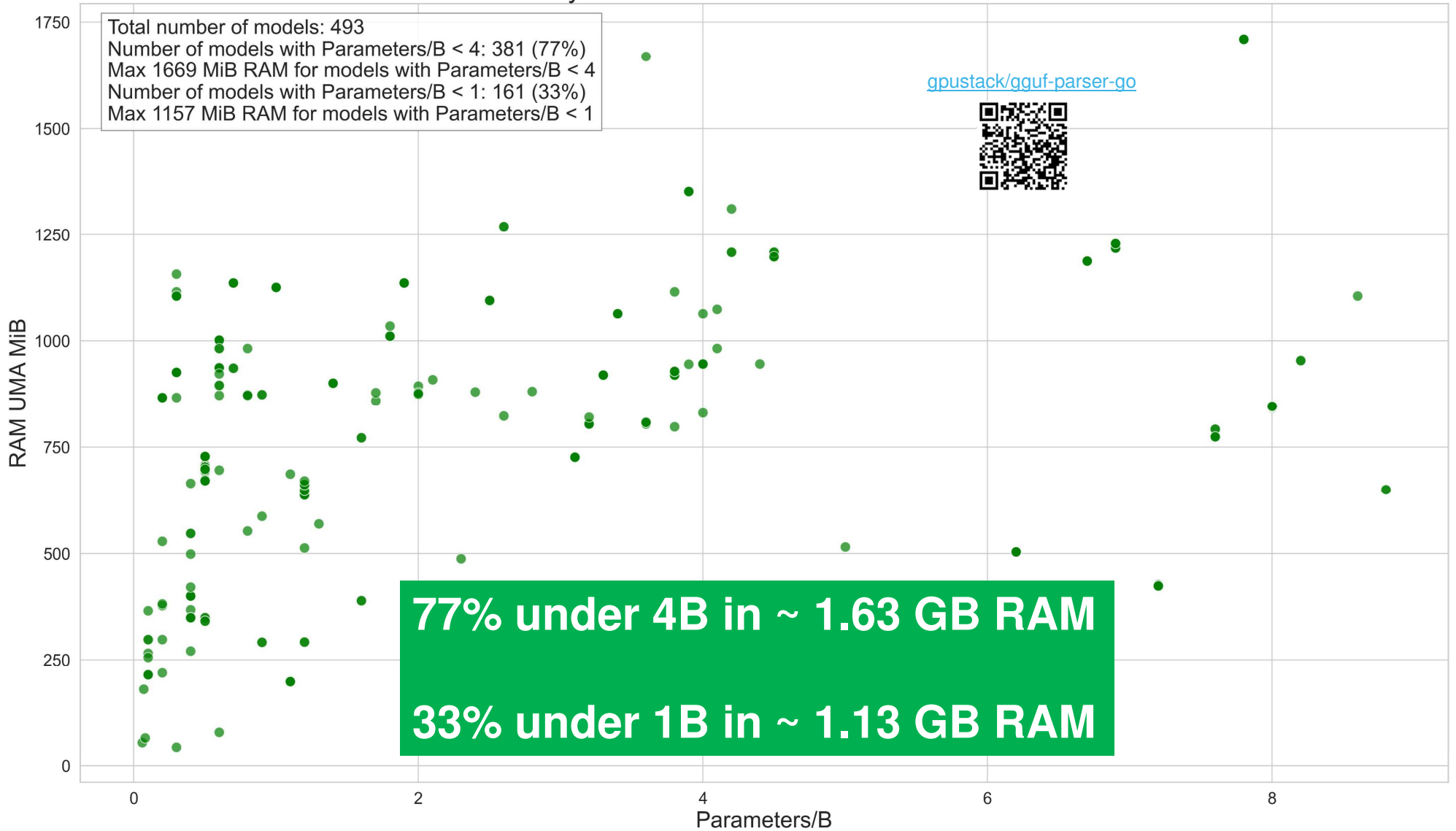
Instruct
849 models

Thinking
3,210 models

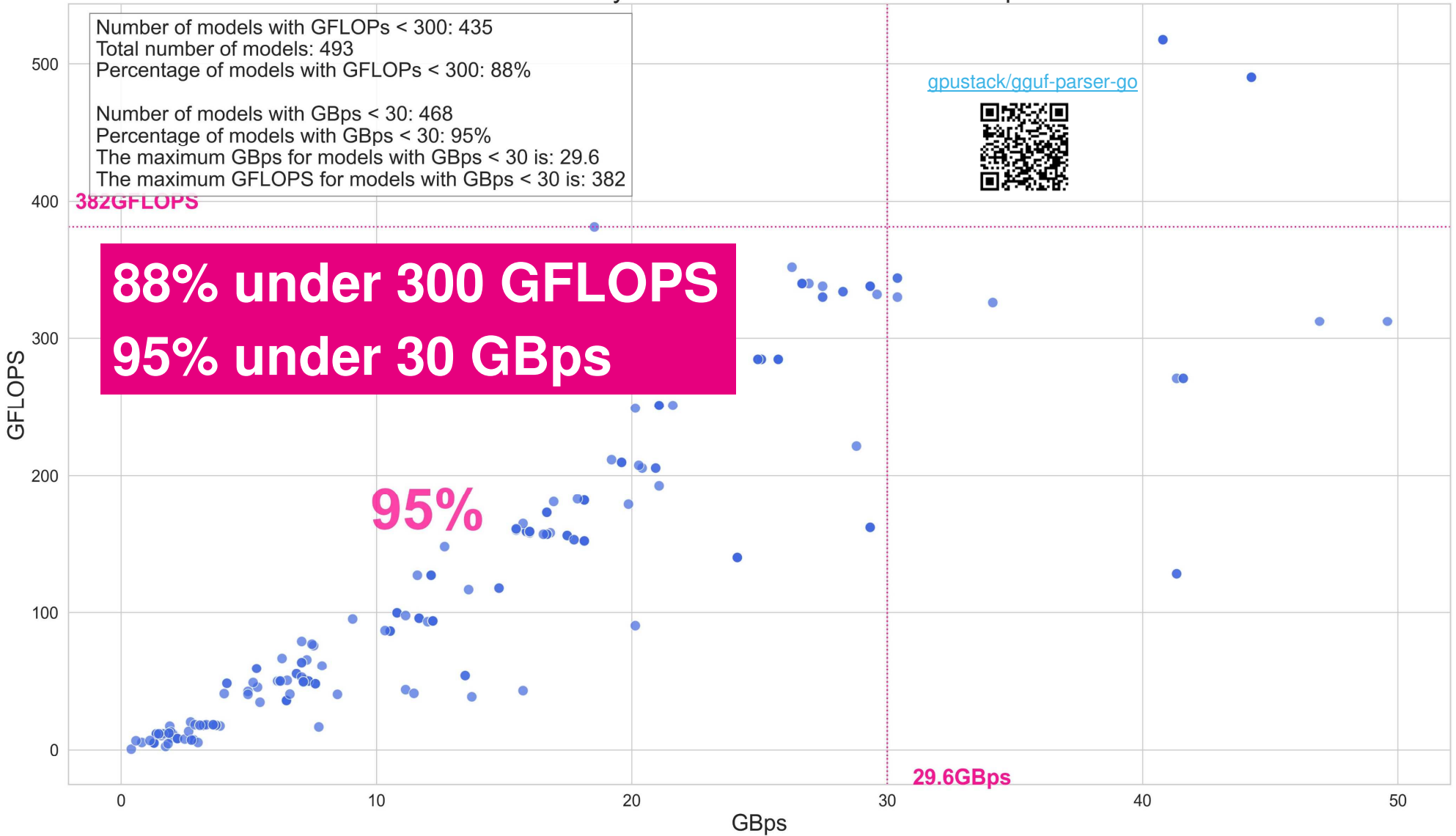


GGUF

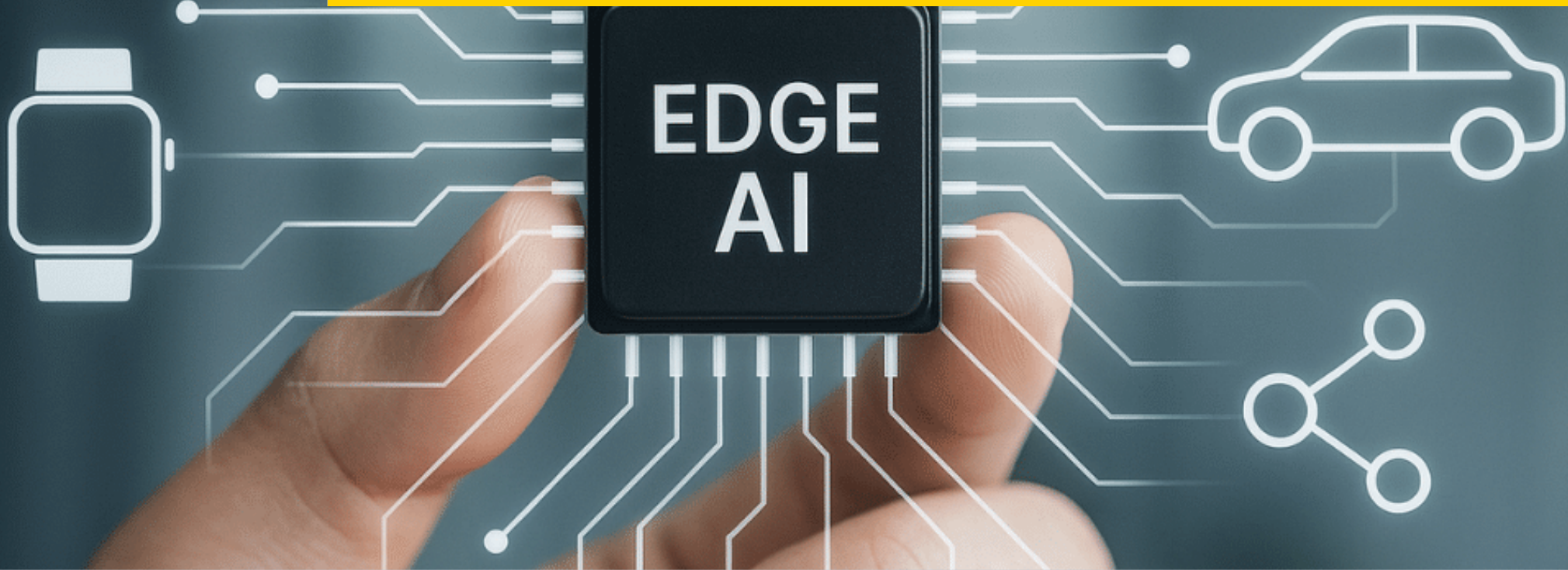
Audio Analysis at 5 token/s: RAM MiB vs Parameters/B



Audio Analysis at 5 token/s: GFLOPS vs GBps



New product categories



Conversational AI vs Industry domain

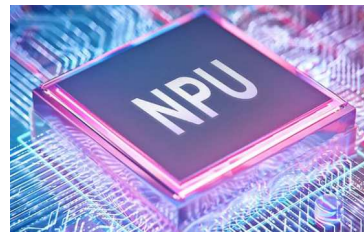
Automotive

Robotics

Consumer & Appliances



Draft demo on Intel SoC



Qwen2.5 (1.5B)
tk/s = 26



Kokoro (82M)
TTS
RTF = 0.194



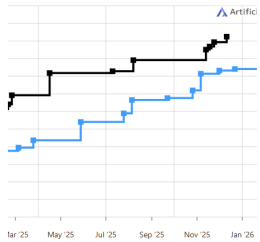
Whisper tiny
(39M) STT
RTF = 0.05
WER = 0.065



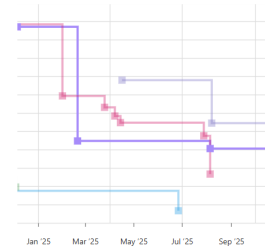
Note: preparation on-going

Conversational AI: further studies



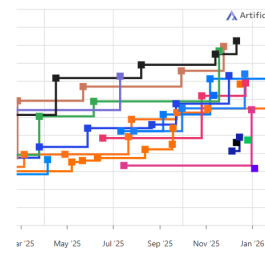


Intense competition between proprietary and open-weight models

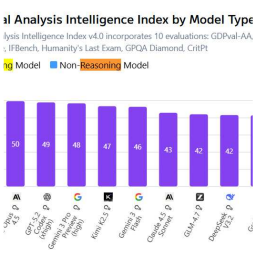
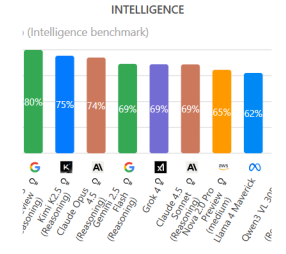


Optimization for price/performance
Audio Latency < 300ms

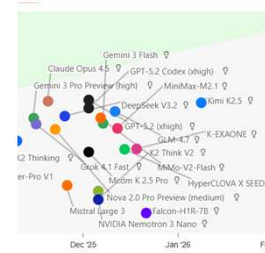
Narrowing the gap in "intelligence" between leading models



Increasing multimodality.

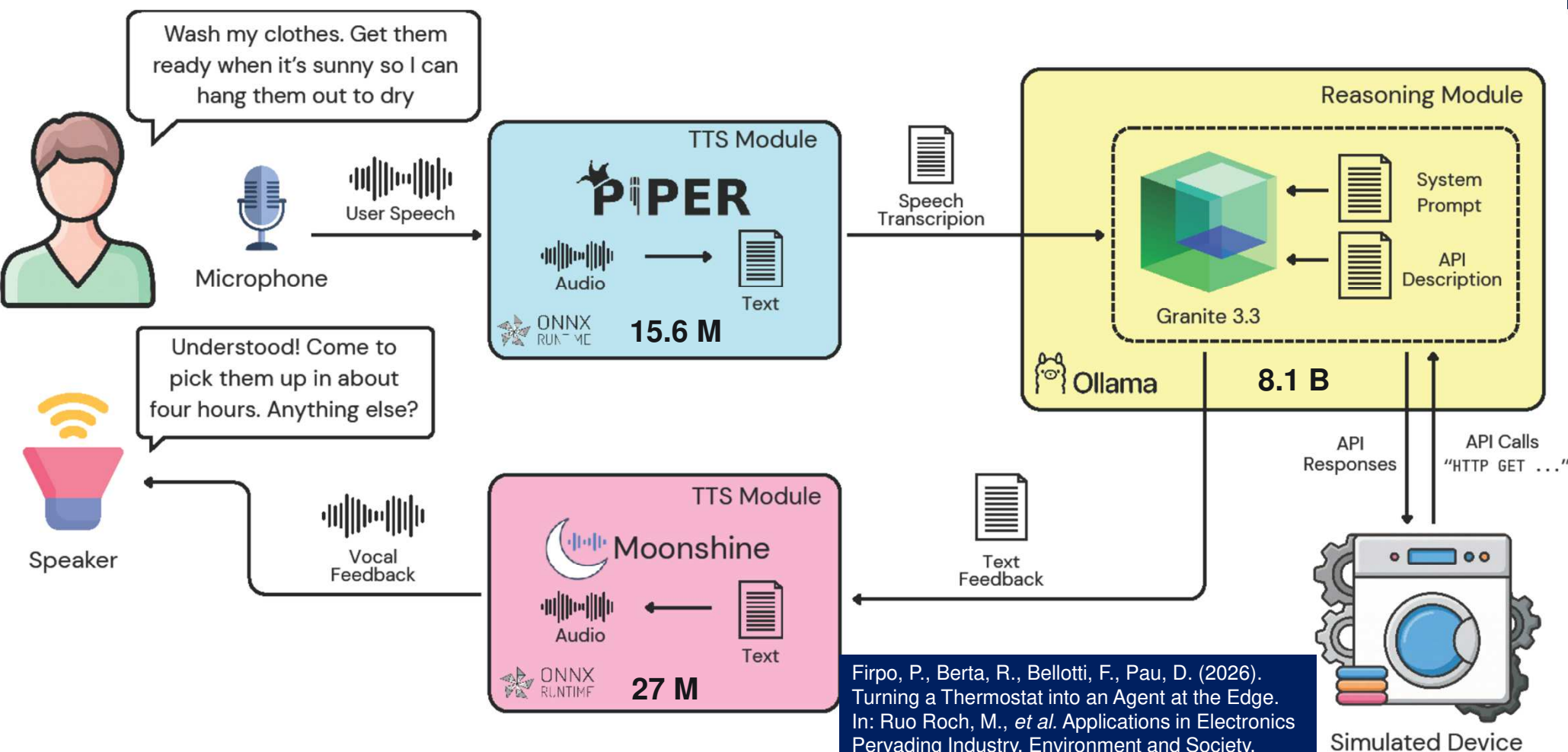


The rise of reasoning models



Agentic capabilities, and expanding context windows





Firpo, P., Berta, R., Bellotti, F., Pau, D. (2026). Turning a Thermostat into an Agent at the Edge. In: Ruo Roch, M., *et al.* Applications in Electronics Pervading Industry, Environment and Society. ApplePies 2025. Lecture Notes in Electrical Engineering, vol 1553. Springer, Cham. https://doi.org/10.1007/978-3-032-17174-0_26



Conversation with thermostat's agent



Firpo, P., Berta, R., Bellotti, F., Pau, D. (2026). Turning a Thermostat into an Agent at the Edge. In: Ruo Roch, M., *et al.* Applications in Electronics Pervading Industry, Environment and Society. ApplePies 2025. Lecture Notes in Electrical Engineering, vol 1553. Springer, Cham. https://doi.org/10.1007/978-3-032-17174-0_26

User: I feel cold in the house. Can you warm the room a bit?

[Internal Action Plan - not spoken aloud]

- 1) Access current room temperature.
- 2) Check weather forecast for location.
- 3) If the current room temperature is below the forecasted average and there are no conflicts, initiate a gradual increase in the setting by 2 °C.
- 4) Set a reminder to recheck in 15 minutes to avoid overshooting.

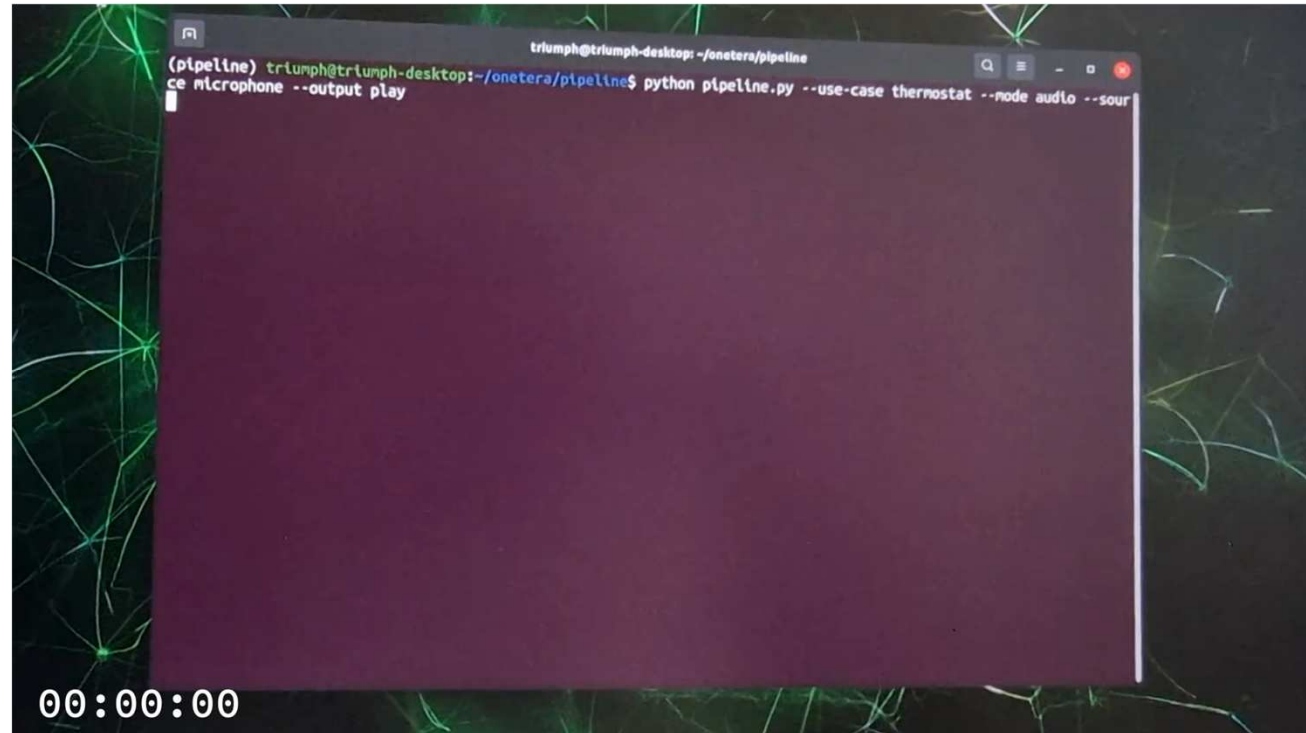
Agent: Your room will start to feel warmer as I slowly increase the temperature by 2 degrees. I aim to reach

a comfortable reference temperature based on today's forecast, ensuring we don't overheat. Would you like to provide any additional preferences or should I proceed with these settings?

User: I'm out in a couple of hours and I'll be back tomorrow morning. I want the home to be warm when I'm back, but I don't need to heat the house when I'm away.

Agent: Understood! I will maintain this temperature for your return, ensuring a cozy environment without unnecessary energy use. Enjoy your time out and rest assured, your home will welcome you with comfortable warmth tomorrow morning.

Conversation with thermostat's agent



Firpo, P., Berta, R., Bellotti, F., Pau, D. (2026). Turning a Thermostat into an Agent at the Edge. In: Ruo Roch, M., *et al.* Applications in Electronics Pervading Industry, Environment and Society. ApplePies 2025. Lecture Notes in Electrical Engineering, vol 1553. Springer, Cham. https://doi.org/10.1007/978-3-032-17174-0_26

Conversation with washing machine's agent



“Deployment and Characterization of a Conversational Generative Agent on Edge Systems”
Journal of Systems Architecture
Under review

User: I need to wash my cotton clothes. Can you schedule the washing machine for me?

[Internal Action Plan - not spoken aloud]

- 1) Select the "Cotton 40 °C" program with standard spin.
- 2) Query the water price API to identify the cheapest tariff in the next 12 hours.
- 3) Query the electricity price API and PV production API to locate the best overlap between low cost and high PV availability.
- 4) Check the weather forecast API to ensure that drying conditions are favorable at cycle completion.
- 5) Schedule the wash to finish shortly before the optimal drying window, while minimizing both water

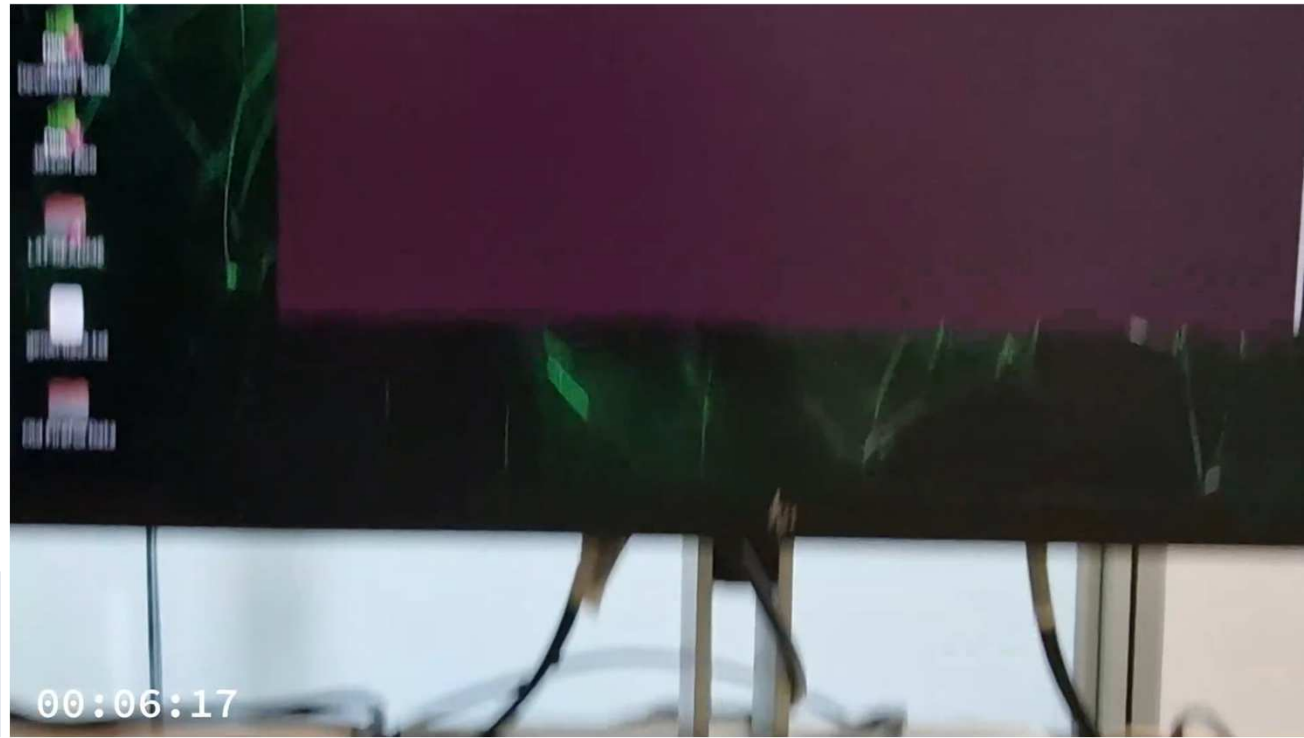
and electricity costs.

Agent: Your cotton load will be washed at 40 °C and ready just before the afternoon sun, so you can hang it out to dry at the lowest cost. Would you like to add details or proceed?

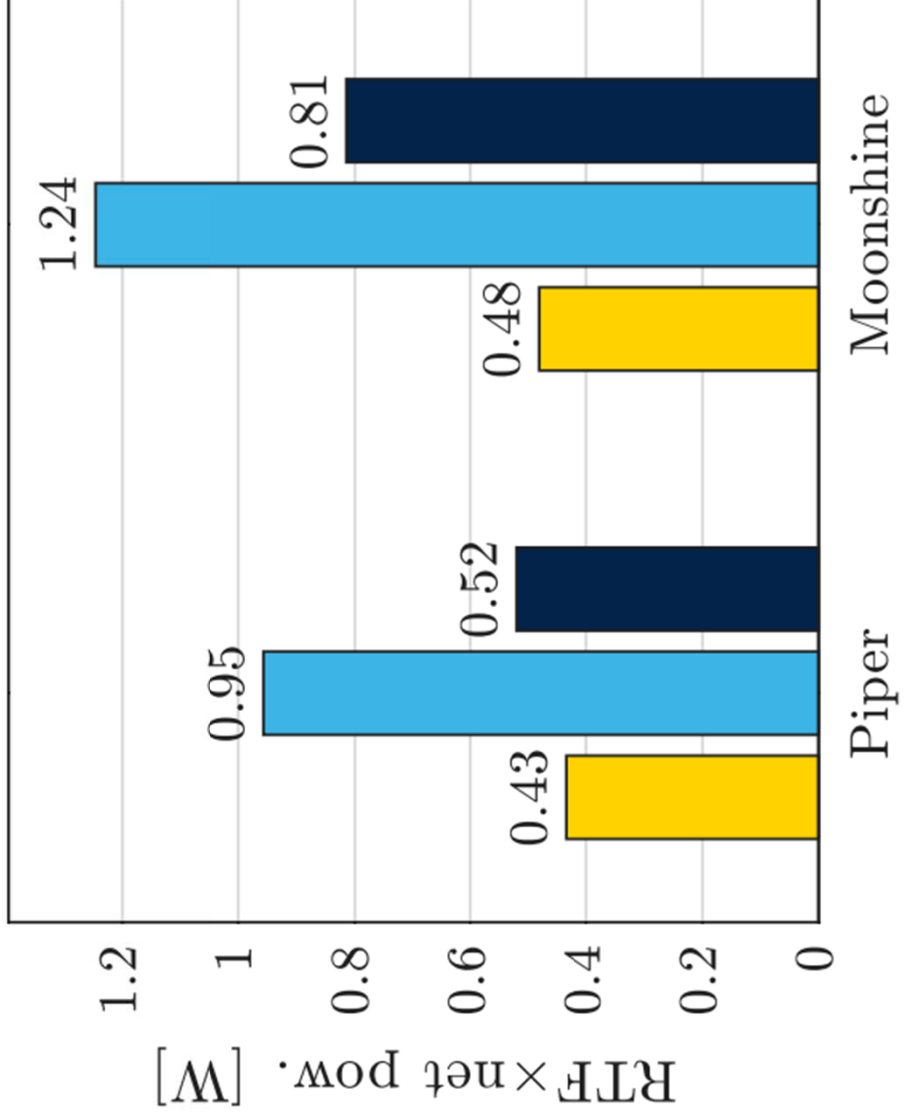
User: I'm leaving home soon and will be back in the evening. I want the laundry done when I return, but I don't want it sitting wet for too long.

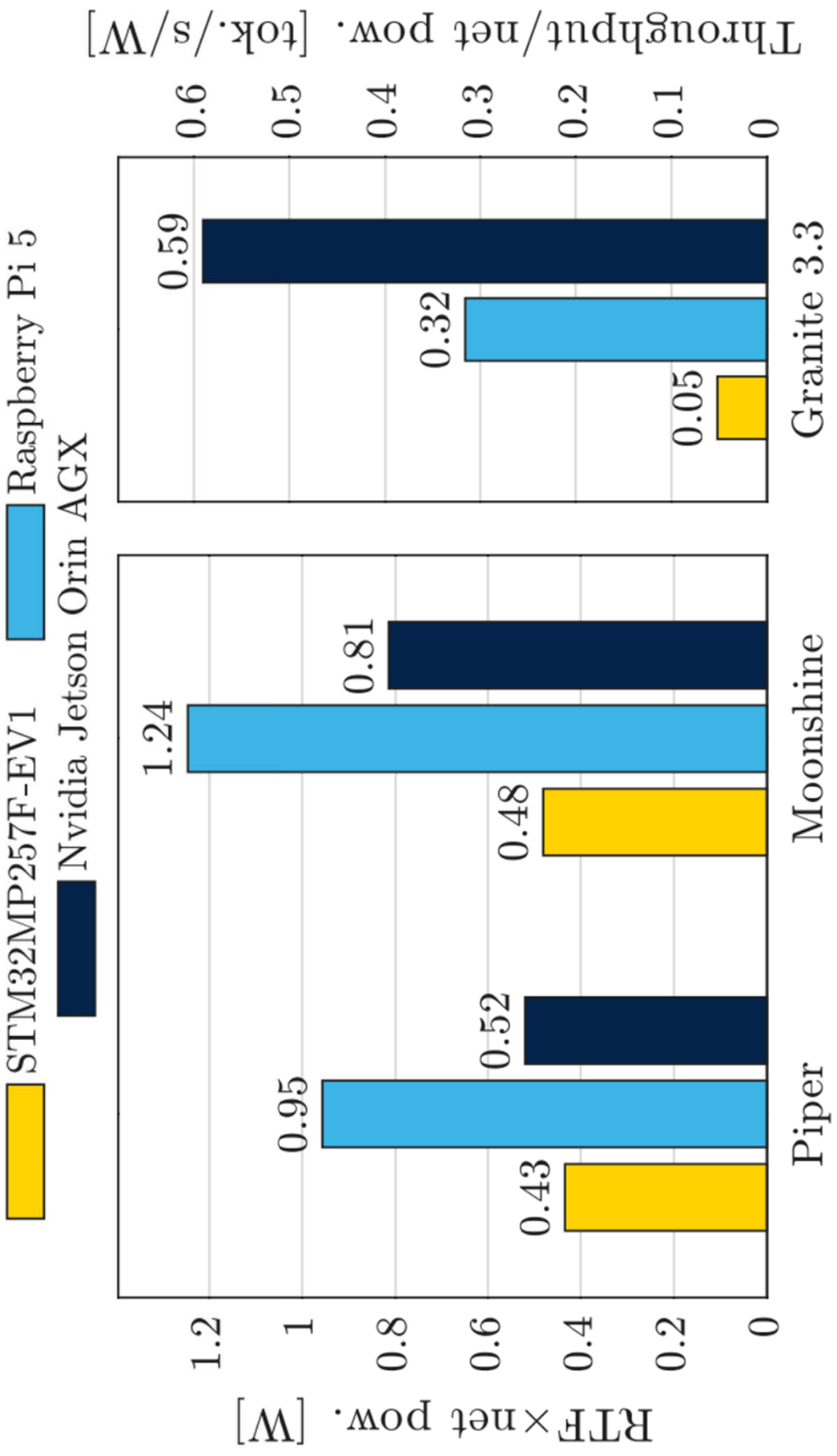
Agent: Understood! I will delay the start so that the cycle finishes right before you return. This way, your clothes will be freshly washed and ready without staying damp in the drum. You can enjoy your day knowing your laundry will be perfectly timed.

Conversation with washing machine's agent

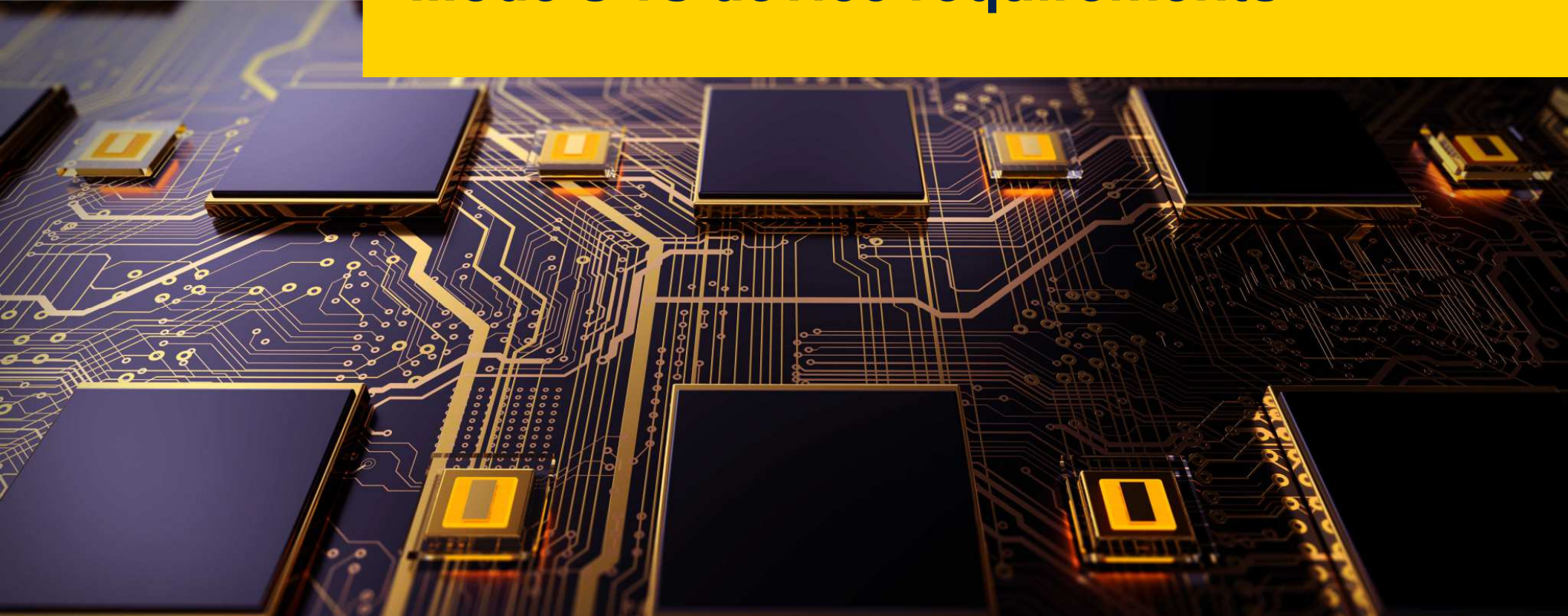


**“Deployment and Characterization of a
Conversational Generative Agent on Edge Systems”**
Journal of Systems Architecture
Under review



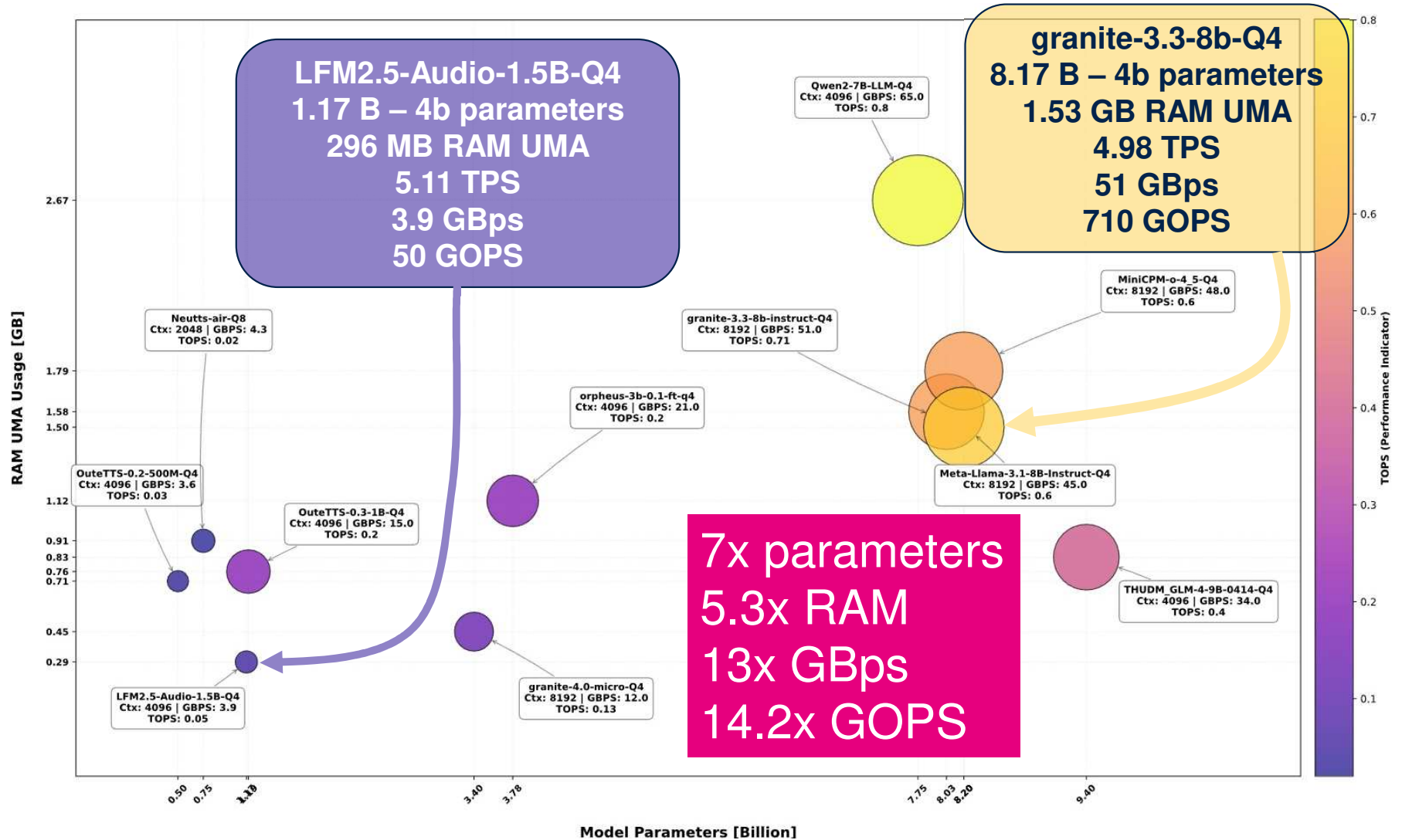


Models vs device requirements



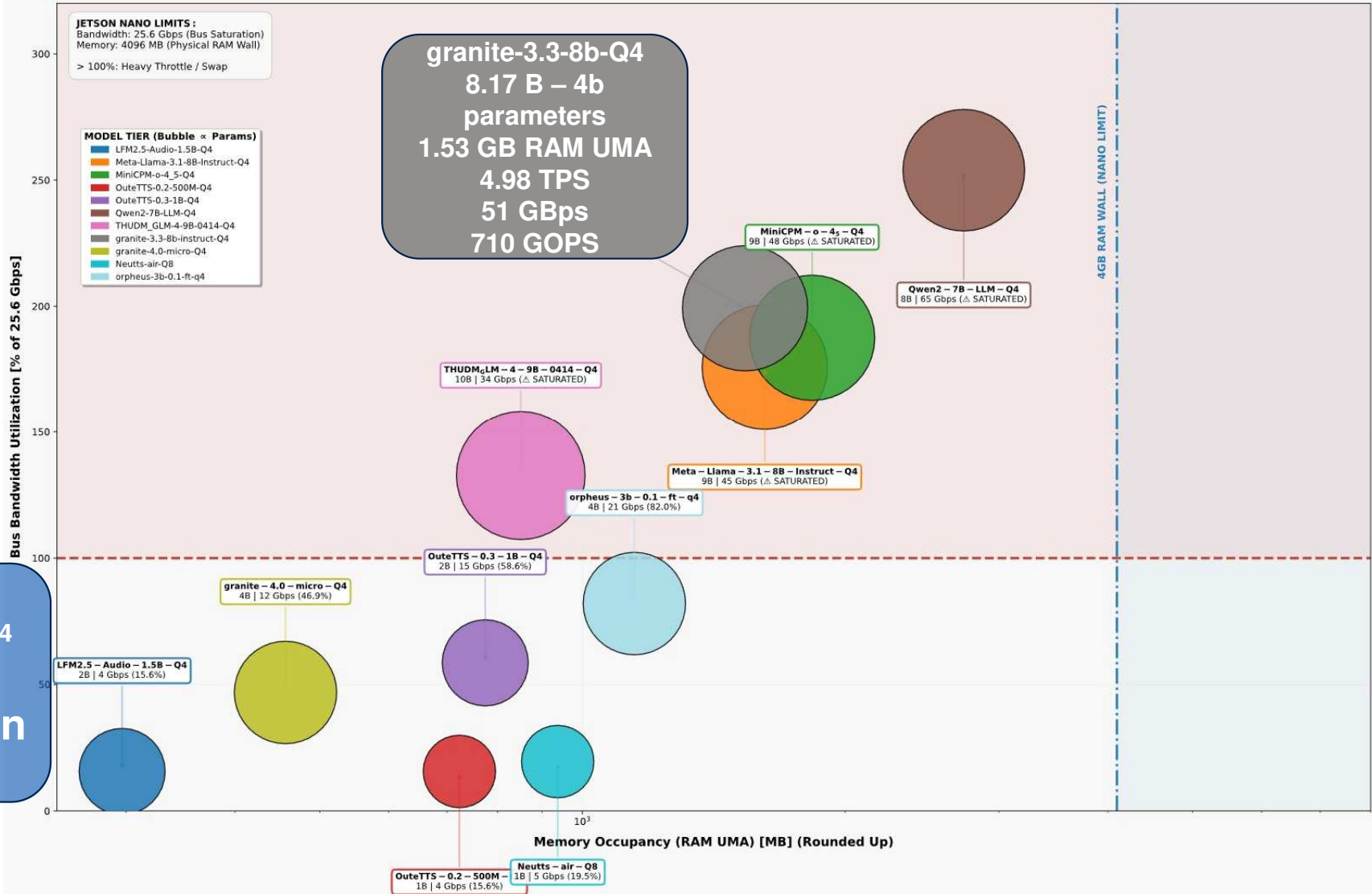


Speech to speech AI Benchmark @ 200ms latency





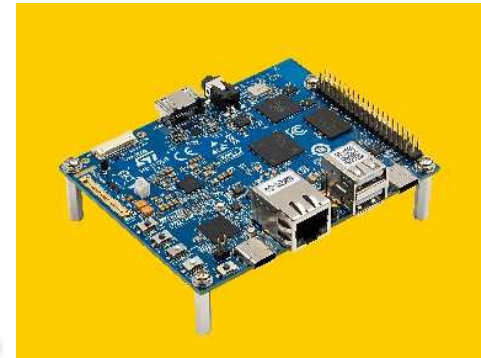
Jetson Nano Performance: Bus & Memory Saturation Analysis



LFM2.5-Audio-1.5B-Q4
15.6% Bus
B/W utilization



Speech 2 Speech: LiquidAI/LFM2.5-Audio-1.5B-GGUF



STM32MP257F-DK

OpenLinux on Dual
Arm A35 1.5 GHz



```
danilo@danilo-HP-Z6-G5-Workstation-Desktop-PC: ~ -- ssh root@192.168.1.14
danilo@stm32mp2-e3-cf-bb:~/genAI/liquid-audio$

top - 11:30:07 up 20:05, 3 users, load average: 0.87, 1.24, 0.85
Tasks: 158 total, 1 running, 157 sleeping, 0 stopped, 0 zombie
%Cpu(s): 0.2 us, 0.5 sy, 0.2 ni, 98.7 id, 0.0 wa, 0.5 hi, 0.0 si, 0.0 st
MiB Mem : 3771.2 total, 741.2 free, 533.4 used, 2695.9 buff/cache
MiB Swap: 32759.3 total, 32749.1 free, 10.2 used, 3237.8 avail Mem

  PID USER      PR  NI   VIRT   RES   SHR  S  %CPU  %MEM    TIME+  COMMAND
    58 root        rt   0     0     0     0   0   S   1.0   0.0   2:33.68 sugov:0
   56319 root       20   0   7488   4480  2560  R   1.0   0.1   0:19.64 top
   1243 root       39  19 132980  82560 26880  S   0.7   2.1  12:25.95 python3
    970 root       39  19 345936  27288 10112  S   0.3   0.7   7:01.27 netdata
   56310 root       20   0   3184   2176  1920  S   0.3   0.1   0:00.48 dropbear
     1 root       20   0  18528  10436  7236  S   0.0   0.3   0:25.84 systemd
     2 root       20   0     0     0     0   S   0.0   0.0   0:00.05 kthreadd
     3 root       20   0     0     0     0   S   0.0   0.0   0:00.00 pool_workqueue_release
     4 root       0 -20   0     0     0   I   0.0   0.0   0:00.00 kworker/R-rcu_g
     5 root       0 -20   0     0     0   I   0.0   0.0   0:00.00 kworker/R-rcu_p
     6 root       0 -20   0     0     0   I   0.0   0.0   0:00.00 kworker/R-slub
     7 root       0 -20   0     0     0   I   0.0   0.0   0:00.00 kworker/R-nets
    12 root       0 -20   0     0     0   I   0.0   0.0   0:00.00 kworker/R-mm_pe
    13 root       20   0     0     0     0   I   0.0   0.0   0:00.00 rcu_tasks_kthread
    14 root       20   0     0     0     0   I   0.0   0.0   0:00.00 rcu_tasks_trace_kthread
    15 root       20   0     0     0     0   S   0.0   0.0   0:02.57 ksoftirqd/0
    16 root       20   0     0     0     0   I   0.0   0.0   0:06.66 rcu_preempt
    17 root       rt   0     0     0     0   S   0.0   0.0   0:00.09 migration/0
    18 root       20   0     0     0     0   S   0.0   0.0   0:00.00 cpuhp/0
    19 root       20   0     0     0     0   S   0.0   0.0   0:00.00 cpuhp/1
```



Liquid

Our technology starts with You

danilo.pau@st.com

Visit us at the ST booth

© STMicroelectronics - All rights reserved.

ST logo is a trademark or a registered trademark of STMicroelectronics International NV or its affiliates in the EU and/or other countries.

For additional information about ST trademarks, please refer to www.st.com/trademarks.

All other product or service names are the property of their respective owners.

