Performance Evaluation of Whisper-Series Speech Transcription Models on Raspberry Pi

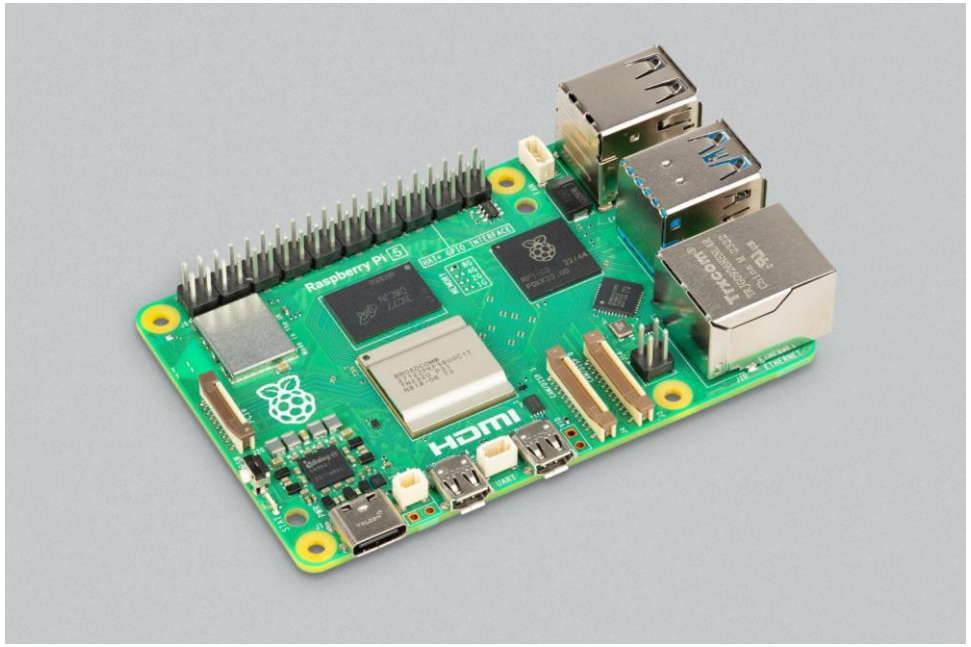
Yue Cao
Syracuse University
SEC ’25 Edge Intelligence Workshop

SEC ’25 — The Tenth ACM/IEEE Symposium on Edge Computing

Motivation & Background 1: Robot Onboard Computer

The Raspberry Pi is a popular low-cost onboard computer for mobile robots.

- Low cost
- Compact size
- Rich I/O



Motivation & Background 1: Robot Onboard Computer

TurtleBot 3

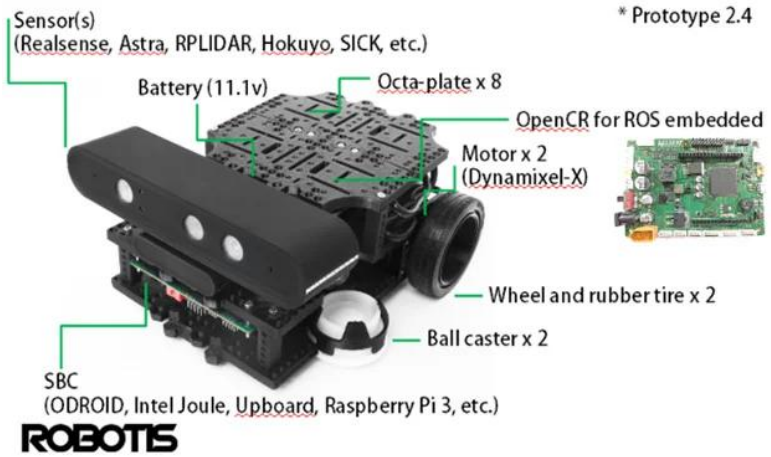


Image: ROBOTIS

Unitree Go1-Edu

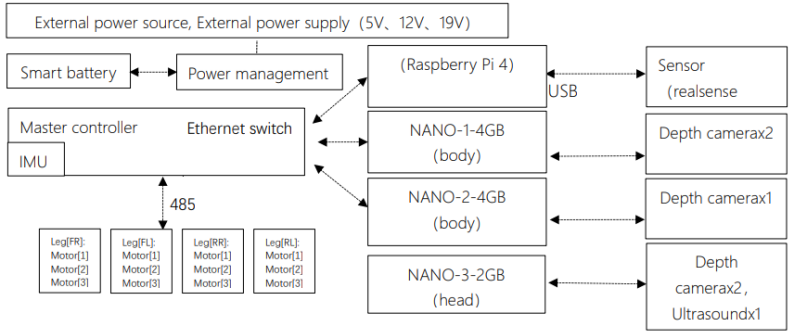


Image: Unitree

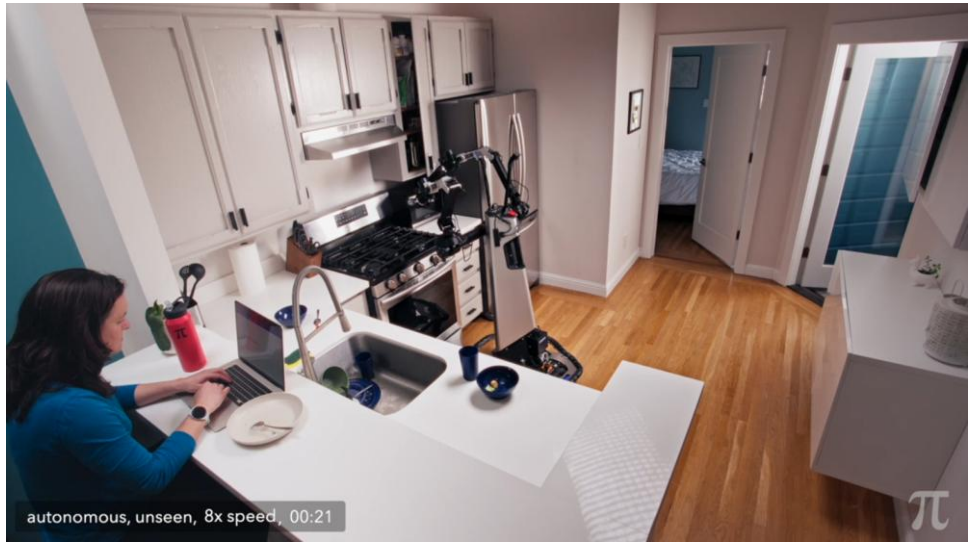
Motivation & Background 2: Large Models in Robotics

The rise of Transformer-based model in AI

- "Attention Is All You Need", 2017
- ChatGPT, 2022

Growing interest in deploying large models in robotics

- Examples: Google RT-2, Physical Intelligence $\pi_{0.5}$



Video Screenshot: Physical Intelligence π_0.5

Motivation & Background 2: Large Models in Robotics

Current pipeline

A better pipeline

- Motivation & Background 1: Robot Onboard Computer
- Motivation & Background 2: Large Models in Robotics
- Overview
- Evaluation Set-up
- Evaluation Results
- Summary and Conclusions

Overview

What to study:

Can speech transcription model operate efficiently on Raspberry Pis?

Overview

What to study:

Can speech transcription models operate efficiently on Raspberry Pis?

Speech Transcription Models I tested:

- **OpenAI Whisper:**
  multilingual speech transcription model, released in 2022

- **Faster-Whisper:**
  a re-implementation of OpenAI's original Whisper models.
  Optimized for resource-constrained settings using CTtranslate2.

Overview

What to study:

Can speech transcription models operate efficiently on Raspberry Pis?

Speech Transcription Models I tested:

Variant	Params	Impl. (Type)	Size
`tiny.en`	39 M	OpenAI (fp32)	151 MB
		Faster-W (fp32)	75.5 MB
		Faster-W (int8)	75.5 MB
`base.en`	74 M	OpenAI (fp32)	290 MB
		Faster-W (fp32)	145 MB
		Faster-W (int8)	145 MB
`distil-small.en`	166 M	Faster-W (fp32)	332 MB
		Faster-W (int8)	332 MB

Overview

What to study:

Can speech transcription models operate efficiently on Raspberry Pis?

Raspberry Pis I tested:

	Processor	Price before Dec25	Price since Dec25
Pi 4 (4GB RAM)	1.5GHz quad-core Cortex-A72	$55	$60
Pi 5 (4GB RAM)	2.4GHz quad-core Cortex-A76	$60	$70
Pi 5 (8GB RAM)	2.4GHz quad-core Cortex-A76	$80	$95
Pi 5 (16GB RAM)	2.4GHz quad-core Cortex-A76	$120	$145

Raspberry Pi Price

Top stories

Yardbarker

Unfortunate Price Hike For Raspberry Pi Computers Coming

4 hours ago

PC Guide

Raspberry Pi 5 and Pi 4 price hikes now in place due to industry-wide memory…

2 days ago

Lowyat.NET

Raspberry Pi The Latest Victim Of AI Craze; Raises Prices On Products

2 days ago

HotHardware

Raspberry Pi Spoiled By Price Hikes But There's A New 1GB Model For $45

2 days ago

More news

NEWS GADGETS TECH

DRAM it! Raspberry Pi raises prices

The 16GB version of the Raspberry Pi 5 (pictured) has jumped from $120 to $145. Image: Raspberry Pi

/ Increasing RAM costs are making Raspberry Pi 4 and 5 models up to 20 percent more expensive.

by Jess Weatherbed
Dec 2, 2025, 7:16 AM EST

10 Comments

Evaluation Set-up

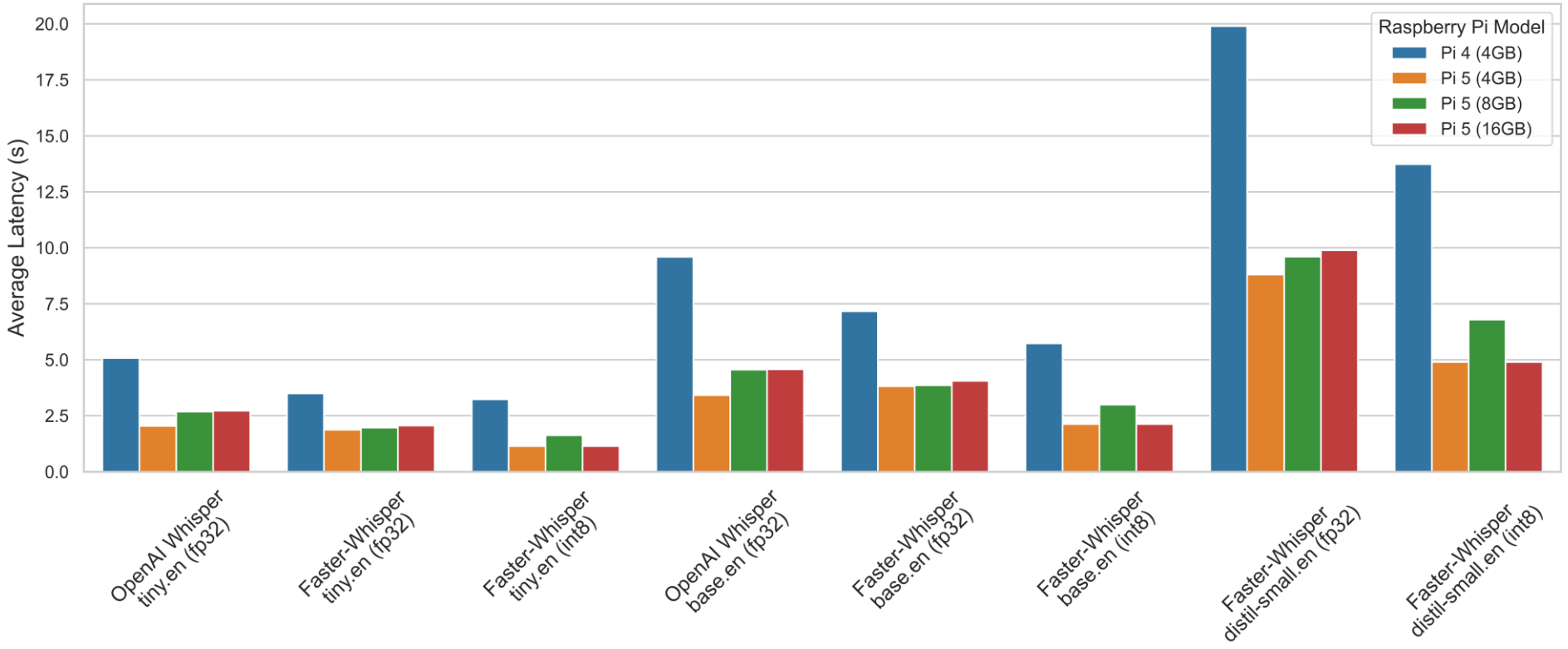
Two Test Sets of .mp3 audio clips were used in the evaluation:

- **Test Set 1: Common Voice**
  10 audio short-text clips from Common Voice dataset.
  Examples:
  “Both the engines and the gearbox proved to be unreliable.”
  “Once the plan is finalized, the implementation phase begins.”

- **Test Set 2: AP News**
  5 long-text from AP News on YouTube.
  Each is about 1-2 minutes long.
  Average duration is 87.8 seconds

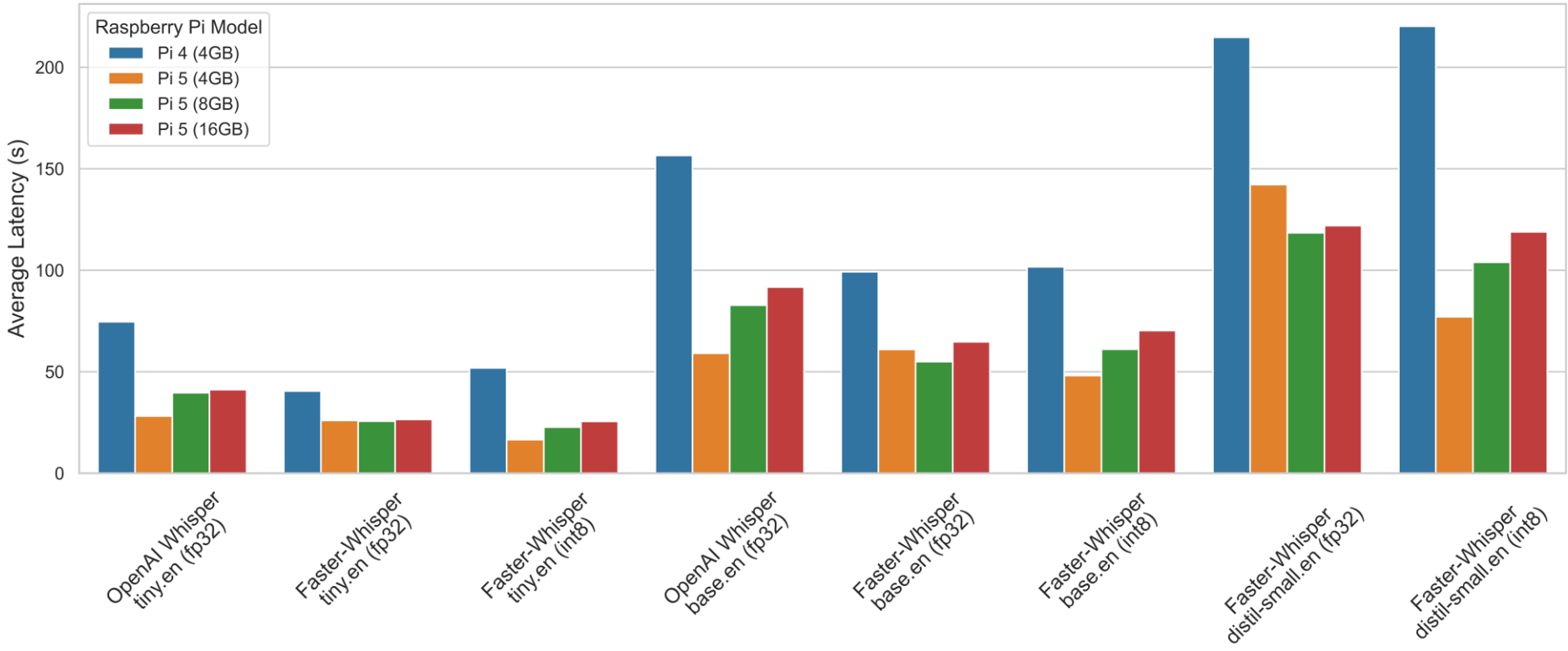
Evaluation Results - Latency

Latency result on Common Voice test set (short text)



Evaluation Results - Latency

Latency result on AP News test set (long text)



Evaluation Results – Transcription Accuracy

Peak RAM usage of Faster-Whisper int8 models on Raspberry Pi 5 (4GB).

Test Set	tiny.en	base.en	distil-small.en
Common Voice	576 MB	879 MB	1522 MB
AP News	745 MB	1304 MB	2076 MB

Evaluation Results – Transcription Accuracy

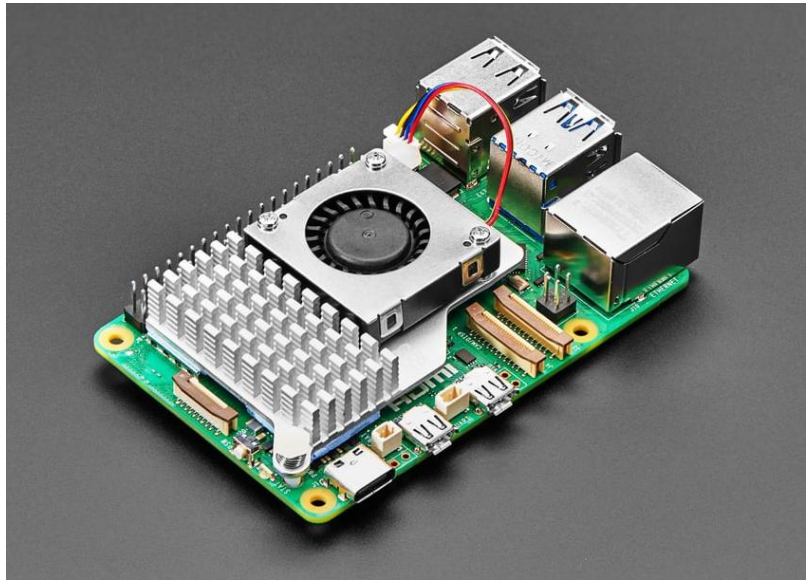
Word Error Rate (WER) results on Raspberry Pi 5 (4GB).
Lower values indicate higher accuracy

Model / Impl.	Common Voice		AP News	
	Mean	Median	Mean	Median
`tiny.en`				
OpenAI Whisper (fp32)	24.5%	4.5%	17.7%	9.4%
Faster-Whisper (fp32)	20.3%	0.0%	16.1%	10.2%
Faster-Whisper (int8)	20.3%	0.0%	14.2%	9.1%
`base.en`				
OpenAI Whisper (fp32)	12.0%	4.2%	16.6%	17.4%
Faster-Whisper (fp32)	10.0%	0.0%	12.2%	10.0%
Faster-Whisper (int8)	10.0%	0.0%	12.9%	9.7%
`distil-small.en`				
Faster-Whisper (fp32)	8.9%	4.2%	60.5%	62.9%
Faster-Whisper (int8)	7.8%	4.2%	62.0%	59.5%

Evaluation Results – Thermal Consideration

Active cooling is necessary for Raspberry Pi 5 to main high performance [1]

Official Raspberry Pi 5 Cooler

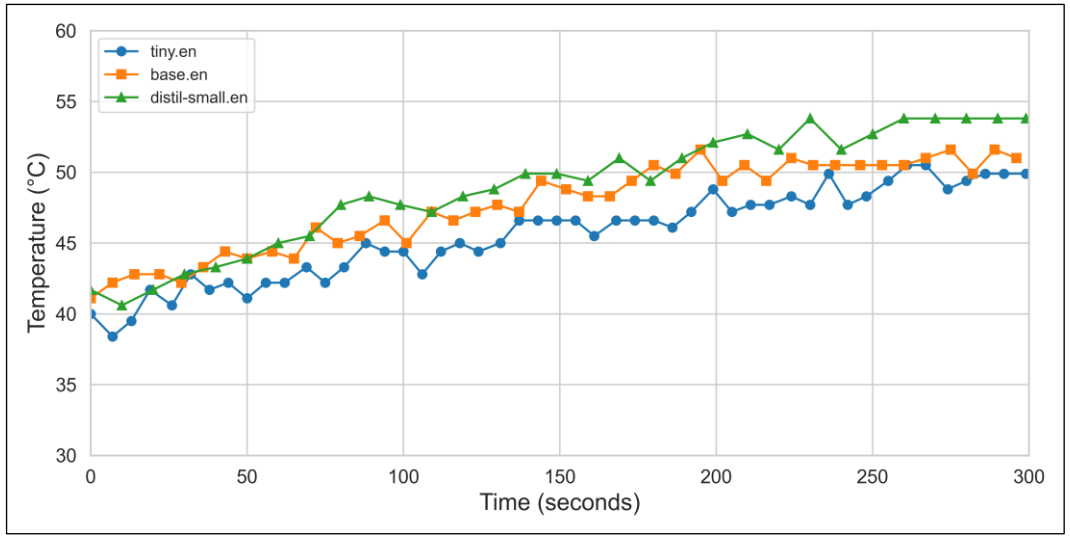


Geekworm Cooler H505 (3rd party)

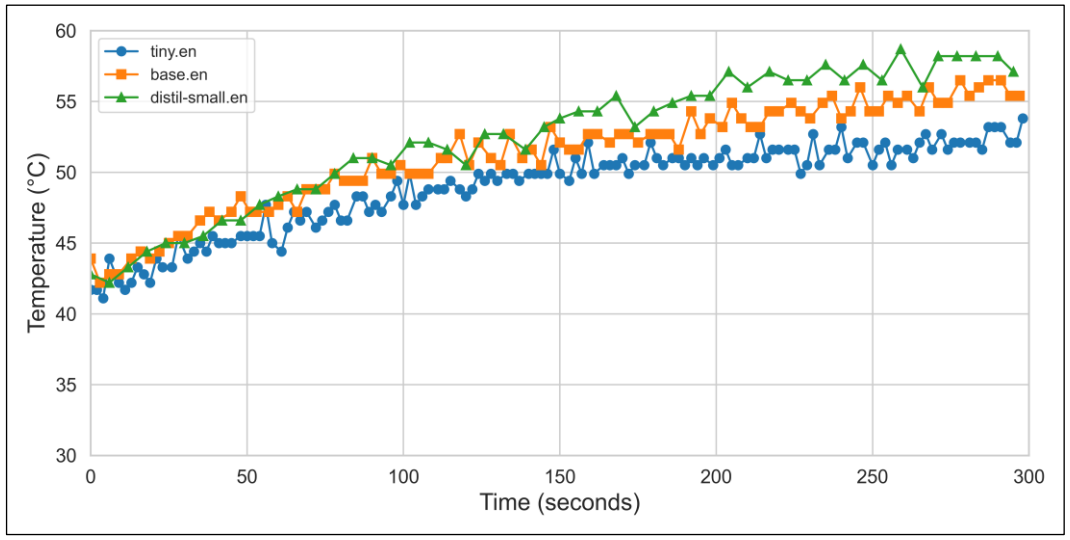
[1] Beserra, David, et al. "Raspberry Pi Single-Board Computers: Cost/Performance Relationship Over Time." IEEE SMC 2024

Evaluation Results – Thermal Consideration

5-second rest interval test.



1-second rest interval test.



Summary and Conclusions

This work evaluated the performance of the Whisper series, on Raspberry Pi 4&5 platforms.

- Raspberry Pi 5 (4GB) offers the best cost-performance ratio.
- Faster-Whisper performs faster than the original OpenAI Whisper on Raspberry Pis
- The tiny.en and base.en models from Faster-Whisper are well-suited Raspberry Pis
- RAM size is not a bottleneck for Whisper model on Raspberry Pis
- Active cooling is necessary

Thank you All!
Welcome to contact me via email or LinkedIn.
I am open to future collaborations.