

YOLO26n · FP32

yolo26n.onnx · graph identity f21c7ca23a529ada... · measured on 3
October 2026

[Download this report as PDF](#) — it is exactly what a client receives.

The verdict

It runs on all 3 configurations measured, with correct results. The fastest:
Intel N100 · ONNX Runtime, at 57.6 ms per inference, about 17 per
second.

RASPBERRY PI 5 ·
ONNX RUNTIME

Runs

126.4 ms

per inference, p50 ·
p95 134.4 ms · 8 per
second

INTEL N100 · ONNX
RUNTIME

Runs

57.6 ms

per inference, p50 ·
p95 58.9 ms · 17 per
second

INTEL N100 ·
OPENVINO

Runs

64.0 ms

per inference, p50 ·
p95 65.6 ms · 16 per
second

Performance

Model inference, in milliseconds. “End to end” adds preparing the input and
processing the output.

CONFIGURATION	P50	P95	P99	END TO END P50	PER SECOND
Raspberry Pi 5 · 8 GB · ONNX Runtime 1.30.0	126.4	134.4	138.3	138.1	8
Intel N100 · 8 GB · ONNX Runtime 1.30.0	57.6	58.9	59.3	67.2	17
Intel N100 · 8 GB · OpenVINO 2026.4.0	64.0	65.6	71.3	72.4	16

Correct results

Outputs are compared with a reference computed with ONNX Runtime on a reference machine, with the same inputs.

Raspberry Pi 5 · ONNX Runtime ✓ finds 100% of the reference objects, with 100% precision, across 12 images

Intel N100 · ONNX Runtime ✓ finds 100% of the reference objects, with 100% precision, across 12 images

Intel N100 · OpenVINO ✓ finds 100% of the reference objects, with 100% precision, across 12 images

What was expected and what happened

Before measuring, the protocol writes down what it expects in each dimension. Then it is checked against what was measured.

DIMENSION	RASPBERRY PI 5 · ONNX RUNTIME	INTEL N100 · ONNX RUNTIME	INTEL N100 · OPENVINO
Valid ONNX model	✓ passes onnx.checker OK	✓ passes onnx.checker OK	—
Runtime load	✓ passes load 57.68 ms	✓ passes load 116.609 ms	✓ passes load 359.373 ms
Operator support	✓ passes session created with no rejected operators	✓ passes session created with no rejected operators	—
Fits in memory	✓ passes peak RSS = 3.0% of RAM	✓ passes peak RSS = 3.3% of RAM	✓ passes peak RSS = 3.5% of RAM
Correct results	✓ passes recall 100% · precision 100% vs reference	✓ passes recall 100% · precision 100% vs reference	✓ passes recall 100% · precision 100% vs reference
p95 ≤ 200 ms	◆ meets it p95 134.3902 ms vs ≤ 200 ms	◆ meets it p95 58.8727 ms vs ≤ 200 ms	◆ meets it p95 65.5797 ms vs ≤ 200 ms
Thermal stability	◆ passes Tmax 65.3 °C	—	—

CPU support	—	—	✓ passes fully executed on CPU
Faster than ONNX Runtime	—	—	◆ no p50 64.045 ms vs 57.5655 ms in C08 (-11.3%)

✓ the expectation was confirmed · ◆ it was unknown and is now measured · · not evaluated in this measurement

14 predictions confirmed, none wrong.

Not evaluated in this measurement: export.

Memory, load and temperature

	RASPBERRY PI 5 · ONNX RUNTIME	INTEL N100 · ONNX RUNTIME	INTEL N100 · OPENVINO
Model load	58 ms	117 ms	359 ms
Peak memory	239 MB · 3.0% of RAM	255 MB · 3.3% of RAM	265 MB · 3.5% of RAM
Temperature	57 → 65 °C	53 → 62 °C	54 → 64 °C
CPU frequency	2400–2400 MHz	2816–2900 MHz	2821–3150 MHz
Thermal throttling	no	no, by frequency (no direct reading)	no, by frequency (no direct reading)

The model

File	yolo26n.onnx · 9.5 MB · FP32
Source	Ultralytics yolo26n.pt (official; never a third-party conversion)
License	AGPL-3.0
Inputs	images FLOAT [1, 3, 640, 640]
Outputs	output0 FLOAT [1, 300, 6]
Graph	ONNX opset 17 · 384 nodes · 23 operator types
File identity	843640bc7ae797dfbf3d9828ced4bfce393a78949b6ed282a0db874f64ef2680
Graph identity	f21c7ca23a529adaf73a497aba4b366aa0d94f9f5bdf02c55aebbbe1460bd90f

Measurement conditions

	RASPBERRY PI 5 · ONNX RUNTIME	INTEL N100 · ONNX RUNTIME	INTEL N100 · OPENVINO
System	Debian GNU/Linux 13 (trixie)	Ubuntu 24.04.5 LTS	Ubuntu 24.04.5 LTS
Python	3.13.5	3.12.3	3.12.3
Effective precision	FP32, the model's own	FP32, the model's own	f32, pinned and verified
Threads	runtime defaults	runtime defaults	runtime defaults
Measurements	100 after 20 warm- up runs	100 after 20 warm- up runs	100 after 20 warm- up runs
Before measuring	device freshly rebooted	device freshly rebooted	device freshly rebooted
Frozen environment	ENV-T1-RPI5-8GB- LAB01-v01	ENV-T2-N100-8GB- LAB01-v02	ENV-T2-N100-8GB- LAB01-v02
Observation	OBS-C07- 20261003T215041Z- b5c74f	OBS-C08- 20261003T224756Z- 1c9208	OBS-C09- 20261003T224811Z- 9534d9

What this report does not cover

- Power draw: it was not measured.
- Integrated GPUs and accelerators: only the CPU was measured.
- More than one input at a time: everything is batch 1, measuring latency.
- Other runtimes, and devices other than the ones shown here.

This is a sample report. The model is the official Ultralytics YOLO26n, which is public, and that is why this report can be published. Client reports are confidential, and the model is deleted when they are delivered.