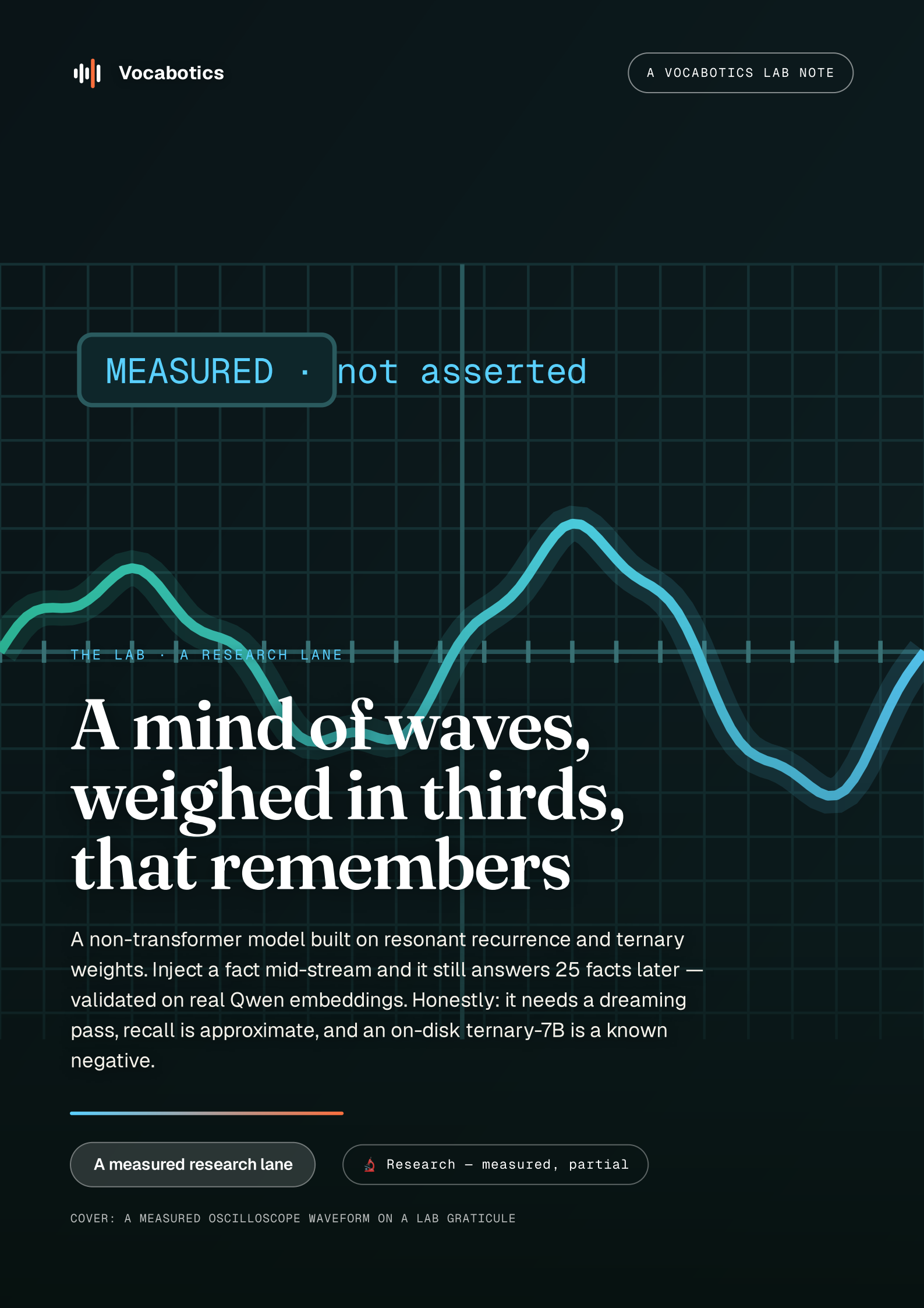




MEASURED · not asserted

THE LAB · A RESEARCH LANE

A mind of waves, weighed in thirds, that remembers

A non-transformer model built on resonant recurrence and ternary weights. Inject a fact mid-stream and it still answers 25 facts later — validated on real Qwen embeddings. Honestly: it needs a dreaming pass, recall is approximate, and an on-disk ternary-7B is a known negative.

A measured research lane

 Research — measured, partial

A mind of waves, weighed in thirds, that remembers

A non-transformer model built on resonant recurrence and ternary weights. Inject a fact mid-stream and it still answers 25 facts later — validated on real Qwen embeddings. Honestly: it needs a dreaming pass, recall is approximate, and an on-disk ternary-7B is a known negative.

01 The claim, in brief

02 Does-it-stick: a genuine, validated GO

03 Reshape beats route

04 The honest negatives — on the page, not in a footnote

05 How it works, briefly

06 Why it ladders back

07 Sources, glossary & the honesty ledger

Every claim carries its tier · nothing faked

PREPARED

2 July 2026

VERIFIED BY

Jon Ross — Founder, Vocabotics —
15 years building safety-critical
systems

HOW TO READ IT

Drafted with AI, verified by a
named human. Sources at the back.

01 The claim, in brief

Inject a fact mid-stream, ask 25 facts later — it still answers, on a model that does no multiplies.

The costly signal

The measured result

Every number carries its own honesty tier — the tiering is the credibility.

GO

does-it-stick, with dreaming

 RESEARCH

cos 0.77

25 facts, drift-bounded

 RESEARCH

cos 0.87

on real Qwen-0.5B embeddings

 RESEARCH

4.4×

saving — 22%-width prefix keeps 95%

 RESEARCH

A transformer remembers by attending over everything it has seen, which is why its cost climbs as context grows. We took a different bet: a **second-order, resonant recurrence** — think of state as a wave evolving under a Verlet-style update — paired with **ternary weights** in `{-1, 0, +1}`, so the core does *additions, not multiplies*. The question that decides whether the bet pays off is blunt: **does a fact stick?**

02 Does-it-stick: a genuine, validated GO

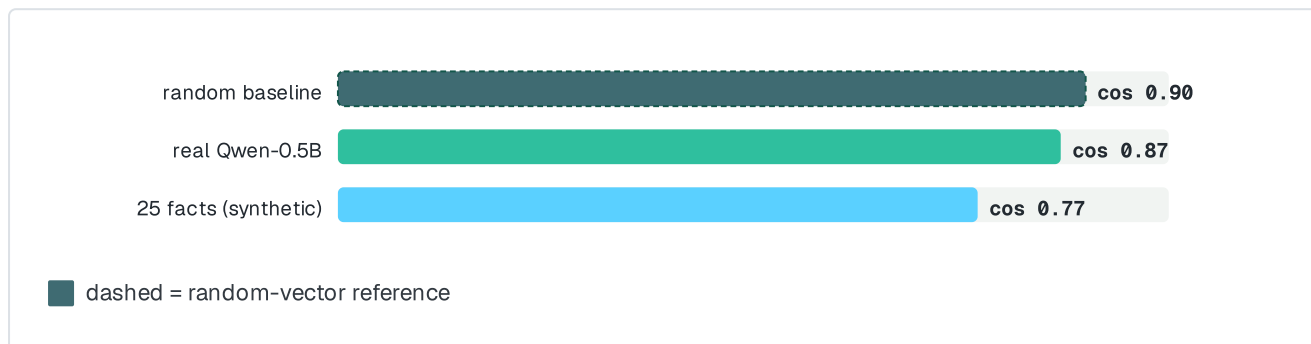
 RESEARCH — MEASURED, PARTIAL

As of 2026-07-02 the plastic-mind go/no-go is **GO** — with caveats we state up front.  Two experiments:

Scale. Inject 25 facts into the stream and query them later: recall holds at **cosine 0.77**, drift-bounded, with capacity around 85–90 (the knee is not reached even at N=60). 📊

On real embeddings. The harder test — running on **real Qwen-0.5B embeddings** rather than synthetic vectors — holds at **cosine 0.87**, versus 0.90 for random. Real, correlated text is only **+0.008 harder**. 📊
No catastrophic forgetting.

That last number is the one we care about most: the memory does not collapse when the inputs are real, correlated language instead of tidy orthogonal vectors.



Does a fact stick? Inject facts mid-stream and query them later. Recall holds at cosine 0.77 across 25 facts; on real Qwen-0.5B embeddings it holds at 0.87 versus 0.90 for random — real, correlated text is only +0.008 harder. Measured, not simulated.

03 Reshape beats route



RESEARCH — MEASURED, PARTIAL

A structural result worth its own line: a **22%-width prefix keeps 95% of accuracy** — a **4.4× saving** — in a regime where a flat network simply collapses. 📊 Reshaping the representation beats routing around it. This is the mechanism that lets a small, cheap model carry more than its parameter count suggests it should.



Reshaping the representation beats routing around it: a 22%-width prefix keeps 95% of accuracy — a 4.4× saving — in a regime where a flat network simply collapses.

04 The honest negatives — on the page, not in a footnote



RESEARCH — MEASURED, PARTIAL

It needs to dream. Without a consolidation ("dreaming") pass, the memory is recency-only — it remembers the last thing, not the right thing. Dreaming is what turns the store into a real associative memory.

Recall is approximate, not verbatim. It answers *about* the fact; it does not quote it back character-for-character. For some uses that is fine; for others it is disqualifying, and we say so.

An on-disk ternary-7B is a known negative. At 7B, on disk, the ternary path runs roughly **7× slower**. We keep this failure visible because a lab that hides its negatives forfeits the right to be believed on its positives.

Why keep a negative on the page?

The ternary-7B result is a real, measured loss. Deleting it would make the wins look cleaner and the whole page less trustworthy. The negative is part of the proof that the positives are honest.



It needs to dream

without a consolidation pass,
memory is recency-only



Approximate recall

it answers about a fact, not
verbatim



Ternary-7B on disk

roughly 7× slower — a measured
loss

The negatives, kept visible — because a lab that hides its losses forfeits the right to be believed on its wins.

05 How it works, briefly

State evolves as a damped, driven oscillation — a wave — so information persists through resonance rather than through an attention matrix. Weights are ternary, so the hot loop is add-and-accumulate, not multiply-accumulate: friendly to cheap silicon and to an in-memory analog substrate later. A separate SSM/wave decode path reaches 45× at 16k tokens with a 1.14 MB state — the kind of number that comes from *not* re-reading the whole context every step.

-1**subtract**

flip and add

0**skip**

no work at all

+1**add**

accumulate

Ternary weights in $\{-1, 0, +1\}$: the hot loop is add-and-accumulate, not multiply-accumulate — friendly to cheap silicon, and to an in-memory analog substrate later.

06 Why it ladders back



A capable model that runs on hardware you already own — because it does additions, not multiplies, and remembers without a datacentre — is access for everyone, not just the people who can rent a cluster.

What is still open — kept visible

The honest edges of this lane, shown next to the wins. None of these is a footnote; each is a claim we have not yet earned — and marking it is what earns the right to be believed on the rest.

- ☐ It needs a 'dreaming' consolidation pass — without it, memory is recency-only.
- ☐ Recall is approximate, not verbatim.
- ☐ An on-disk ternary-7B is an honest negative: roughly 7× slower. We keep it on the page.

07 Sources, glossary & the ledger

Where every claim comes from — and what the plain words mean.

Sources

- 01** **Vocabotics Dashboard — Organ 2 (WaveTernary / The Body), measured 2026-07-02** — Vocabotics internal record (as of July 2026)
- 02** **BitNet: 1.58-bit ternary weights for large language models** — the public ternary-weight line of work our body draws on (as of 2024)

THE HONESTY LEDGER

- Shipped — measured, gate-clean.
- 🔬 Research — measured, partial; the gap is shown.
- 🔭 Vision — designed, honest; not yet built.

The bigger claim goes up a tier, never down. We show the eval beside the demo — the tiering is the credibility.

See it run fast, on hardware you own

The wave-ternary body needs an engine. Read how a real Qwen generates text in QUANTA at 151.7 tok/s on one 3090, token-for-token with llama.cpp.

[/research/gpu-inference-training](#)

Nothing faked.

Measured, not asserted. Every claim carries its own honesty tier and the open edges are shown next to the wins — the tiering is the credibility.

