Vocabotics

A VOCABOTICS LAB NOTE

THE LAB · A RESEARCH LANE

The calculated mind: it reasons or refuses

A mind where meaning is calculated, not guessed, and only the voice is learned. Confabulation drops from 87% to 0.0% at the type level. It cites from a 0.70 GB store. And — honestly — a cold stranger still scores it 3/10, and 'calculated beats trained at language-model scale' is still open.

A measured research lane

Research — measured, partial

COVER: A MEASURED OSCILLOSCOPE WAVEFORM ON A LAB GRATICULE

Vocabotics

A VOCABOTICS LAB NOTE

The calculated mind: it reasons or refuses

A mind where meaning is calculated, not guessed, and only the voice is learned. Confabulation drops from 87% to 0.0% at the type level. It cites from a 0.70 GB store. And — honestly — a cold stranger still scores it 3/10, and 'calculated beats trained at language-model scale' is still open.

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

> *Watch it refuse to make something up — then cite Canberra from a store that fits on a phone. The refusal is the costly signal.*
>
> The costly signal

The measured result

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

87% → 0.0%

confabulation (type-level)

🟢 SHIPPED

20 GB → 0.70 GB

22.2M citeable facts

🟢 SHIPPED

2 → 3 / 10

cold-judge stranger score

🔬 RESEARCH

0.84

loss — reconstruction, not prediction

🔬 RESEARCH

The industry's default failure mode is confident invention: an AI states a fake citation with the same fluency as a true one. Our wager is architectural. If *meaning* is calculated — reasoned over a small set of semantic primitives — rather than sampled from a distribution, then an answer can be **reasoned or refused**, never bluffed. Only the surface, the voice, is learned.

This is the lane where we are most careful to show the eval beside the demo, because the strongest results and the most honest gaps sit right next to each other.

02 No-bluff, at the type level

🟢 SHIPPED — MEASURED, GATE-CLEAN

The measured result: **confabulation from 87% to 0.0%**, enforced at the type level rather than coaxed with a prompt. 🟢 The mind passes a **no-bluff floor** under a 16-turn adversary that actively tries to make it invent. A provenance invariant runs underneath: **every word traces back to a meaning** it can name.

One consequence is small and telling: the voice layer **can never invert a negation**. "Cannot" cannot silently become "can" — proven on the real wave, not asserted. In a safety context, a flipped negation is the difference between "safe to proceed" and "not safe to proceed", so this is a property, not a nicety.

87%	→ the no-bluff floor	0.0%
confabulation, baseline		type-level, adversarial

Confabulation enforced to zero at the type level — under a 16-turn adversary that actively tries to make it invent. Not coaxed with a prompt: a provenance invariant means every word traces back to a meaning it can name.

03 It cites, from a store that fits on a phone

🟢 SHIPPED — MEASURED, GATE-CLEAN

Knowledge is compressed hard and kept citeable: **60M facts distilled to 22.2M citeable ones**, and **20 GB down to 0.70 GB** — about 28.6 bytes per fact — with provenance intact. 🟢 Ask it for the capital of Australia and it cites Canberra; ask about Austen, da Vinci, Tokyo, and it answers *with the source*, from a store small enough to live on a laptop.

The crisis floor is handled by meaning, not keywords: it catches "I don't want to be here" and suppresses an essay on the word "perish" — the kind of literal misread that keyword systems make.

20 GB	→ ≈ 28.6 bytes / fact	0.70 GB
raw knowledge		22.2M citeable facts

Knowledge compressed hard and kept citeable — about 28.6 bytes per fact, provenance intact, small enough to live on a laptop.

04 The honest gaps — kept visible

RESEARCH — MEASURED, PARTIAL

Here is where a lesser page would go quiet. We won't.

A cold stranger scores it 3/10. Up from 2/10 (knowledge went 0→5, the crisis response went from FAIL to trustworthy) — but capped at 3 by deflection, noun-grab, a garbled DERIVE step and register misreads. It is not yet a pleasant conversationalist to a first-time user.

0.84 is reconstruction, not prediction. The mind reaches loss 0.84 on a single 3090 at 21 ms per step and 377 MiB of VRAM — a genuinely efficient, pure stack — but that number measures reconstruction, not next-token prediction. It does not, on its own, prove the mind is better than a trained one.

"Calculated beats trained at language-model scale" is OPEN. This is the real research question, and we mark it open. Closing it needs downstream evaluations, not a lower reconstruction loss.

Why we publish the 3/10

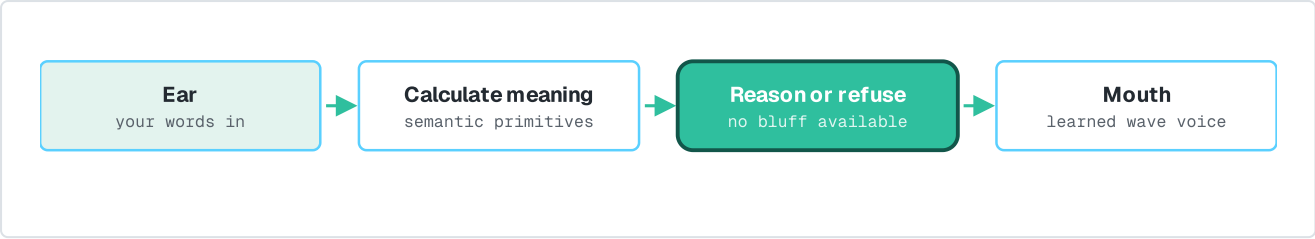
Every other lab shows a demo and hides the eval. Showing the cold-judge score next to the 0→5 knowledge win is what earns the right to be believed on the 0.0% confabulation. Honesty here is not the tax on the claim; it is what makes the claim land.

2 / 10	knowledge 0 → 5	3 / 10
prior cold-judge score		now — honestly, still capped

The gap we keep on the page: a cold stranger scores it 3/10. Up from 2 (knowledge went 0→5, crisis response FAIL→trustworthy) — but capped by deflection, noun-grab and register misreads. Not yet a pleasant conversationalist.

05 How it works, briefly

Meaning is composed over a set of semantic primitives with no learned weights; the wave layer renders *form and feeling*, while a grammar layer gold-fills *reference* (so it says "Tokyo", not a plausible-sounding wrong noun). Six reply moves speak *from* the composed meaning rather than predicting the next token — which is exactly why a refusal is available as a first-class outcome.



Meaning is calculated over semantic primitives, then spoken through a learned wave voice — so 'reason or refuse' is a first-class outcome and the voice can never invert a negation.

06 Why it ladders back

A mind that reasons or refuses is a mind you can trust — and trust is the thing that turns a capable tool into an empowering one. We would rather ship a 3/10 stranger score that never lies than a 9/10 charmer that sometimes does.

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.

- [ ] The cold-judge stranger score is 3/10, not higher — capped by deflection, noun-grab and register misreads.
- [ ] 0.84 is a reconstruction loss, not next-token prediction: the moat claim is not yet earned by it.
- [ ] 'Calculated beats trained at language-model scale' remains OPEN — it needs downstream evals to close.

07 Sources, glossary & the ledger

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

Sources

01 **Vocabotics Dashboard — Organ 3 (The Calculated Mind), measured 2026-07-02** — Vocabotics internal record (as of July 2026)

02 **On measuring hallucination / confabulation in language models** — Vocabotics research note — the no-bluff floor and 16-turn adversary protocol (as of July 2026)

Vocabotics

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 the model this mind speaks through

The calculated mind renders language through a wave model that remembers. Read how a mind of waves, weighed in thirds, holds a fact you inject mid-stream.

/research/wave-ternary-models

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.

Vocabotics · calm, clear, candid · no client names · global rail & metro heritage