LAB NOTEBOOK · dated

THE LAB NOTEBOOK

GLYPH — compressing meaning, not text

GLYPH treats language as a compression of meaning: natural language stripped to ~1,024 atomic primitives and ~10,240 opcodes in a Polish-prefix tree, with a separate channel for tone. Bidirectional NL↔AST↔NL, ~57% smaller than token streams, trainable in ~13 GPU-days. A falsification-first spec — with hard kill-thresholds named up front — offered as an open standard.

The dated R&D archive

Research — measured, partial

COVER: A DATED NOTEBOOK GRATICULE WITH A PLOTTED, RISING MEASURED TRACE — THE LEAD OVER TIME

Vocabotics

A VOCABOTICS LAB REPORT

GLYPH — compressing meaning, not text

GLYPH treats language as a compression of meaning: natural language stripped to ~1,024 atomic primitives and ~10,240 opcodes in a Polish-prefix tree, with a separate channel for tone. Bidirectional NL↔AST↔NL, ~57% smaller than token streams, trainable in ~13 GPU-days. A falsification-first spec — with hard kill-thresholds named up front — offered as an open standard.

PREPARED
25 April 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 entry, dated

The entry, dated

Apr 2026
entry date

Language
category

~1-2 years ahead
lead over the world

🔓 Public
access

What the world had at the time

Everyone tokenises text into fragments and trains ever-larger models on the stream. Treating language instead as a compressible representation of meaning — an algebra you can reason over — is a different bet from the one the whole industry is making.

> *What if you compressed the meaning instead of the text? A dense language of 1,024 primitives that round-trips to English and back — and trains in under two GPU-weeks.*
>
> The costly signal

The measured result

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

1,024
atomic semantic primitives
🔬 RESEARCH

~57%
compression vs token streams
🔬 RESEARCH

~13
GPU-days to train a ~120M model
🔬 RESEARCH

F1-F4
hard kill-thresholds named up front
🔬 RESEARCH

The whole field's default is to chop text into tokens and train larger models on the stream. GLYPH takes a different bet: that **language is a compression of meaning**, and if you compress the *meaning* directly you get something smaller, cheaper to train, and possible to reason over algebraically. This is offered as an **open standard** — the spec and reference implementation are the contribution.

02 What it was

A **super-dense intermediate language for meaning**. Natural language is stripped to roughly **1,024 atomic primitives** and **~10,240 opcodes** arranged in a **Polish-prefix abstract syntax tree**, with a separate **16-dimensional pragmatic channel** carrying tone. Crucially it is **bidirectional**: English → AST → English, so you can compress to meaning and reconstruct the prose.

03 What we built

A specification and the beginnings of a model:

The primitive set, opcode table, and the Polish-prefix AST format.

A **falsification-first** design: hard kill-thresholds **F1–F4** named *before* building — for example, **≥99.7% fact recall** and **≥80% tone** on round-trip — so the project could be *disproven*, not just admired.

An estimate that a **~120M-parameter** model over this representation is trainable in about **13 GPU-days**.

🔬 RESEARCH — MEASURED, PARTIAL

The measured/estimated facts: roughly **57% compression** versus a raw token stream, at a training cost small enough to be within reach of a single machine. The interesting property is not just the size — it is that a meaning-first representation is something you can **reason over**, not merely predict.

04 What we learned — including the honest edges

A spec is not a validated model. The gate is in flight; the falsification thresholds are *named*, not all *cleared*. Those recall and tone bars are targets to beat, and until they are beaten the compression number

is the proven part and the fidelity is the open part.

Naming your own kill-switch is the point. Declaring F1–F4 up front means the project can honestly fail. That is the discipline that makes the wins believable elsewhere in the lab.

Dormant-by-design is an honest status. It was placed on a council-approved deferral — parked deliberately, with the reason recorded, not quietly dropped.

Why open the spec?

A language only becomes a standard if others can adopt it. Opening the GLYPH spec and reference implementation is how you earn adoption and credibility — while the premium tooling stays closed. Open the spec; keep the engine.

05 Where it went / status

Dormant by design, offered openly as a spec. It is the sibling, for *meaning*, of the lab's work on a language that proves *code* — two halves of one idea: that language is something a machine should reason over and verify, not just read and imitate.

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 is a spec with a gate in flight, not a finished, fully-validated model — falsification thresholds are named but not all cleared.
- ☐ Round-trip fidelity (fact recall, tone) has hard pass bars (≥99.7% / ≥80%) that must be met — they are targets to beat, not claims already made.
- ☐ It was placed dormant by design (a council-approved deferral), which is an honest status, not a quiet abandonment.

06 Sources, glossary & the ledger

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

Sources

01 **Vocabotics project audit — GLYPH (dense semantic language, ~57% compression), Apr 2026** — Vocabotics internal project history (as of April 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 language that proves what it emits

GLYPH compresses meaning; QUANTA proves code. Two halves of treating language as something a machine reasons over, not just reads.

/research/quanta

Nothing faked.

A real, dated lab report from our own R&D. Measured, not asserted; the honest negatives are kept on the page next to the wins — because a lab that hides its failures forfeits the right to be believed on its results.

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