METHOD

Know Thyself

A schema for personal memory in LLM conversations

After enough conversations with a model, its memory of you starts to look like a list.

A list isn't bad. Lists are what memory looks like when it has no shape. But the list has a specific failure mode that goes invisible fast: a claim stated once feels the same as a claim stated five times, which feels the same as a claim grounded in five independent events. The distinction between *I said this repeatedly* and *this has been independently confirmed* collapses to nothing. After a year, the model believes things about you that rest on a single inference it made early on and has politely restated back to you ever since.

The fix is not more memory. It is shape.

Consider two claims. The first: *Alex moved to a new city last August, and the first three months were hard — working long hours, her child struggling at school, the running routine she had kept for years stopped.* Specific. Dated. A thing that happened.

The second: *Alex's physical routine is load-bearing for everything else.* Interpretive. A theory. Something Alex and the model have said back and forth for six months.

A flat list puts these on the same footing — one line each in the memory store. But the interpretation is cheap. You can produce interpretations all day without

new evidence. The episode is the thing the interpretation rests on. If the episode is miscoded, the interpretation crumbles. If the interpretation is wrong, the episode remains.

A shaped memory makes the distinction structural.

Reference is what's biographical and verifiable. Where you were born. Who you married. What you do.

Observation is what happened. The three hard months after the move. The first Sunday at the new running group. The conflict with the senior colleague in March. Each one dated, each one contained.

Overlap is a pattern grounded in two or more **independent** observations. Not the same claim restated — the same *shape* recurring across different events. Alex's running routine stopped, then work extended and her child struggled. Her running routine restarted, then work stabilized and her child's grades recovered. Two episodes, same shape. That's an overlap. Her saying "routine matters to me" five times is not.

Those three carry the weight. Four more handle the edges.

Novel is an interpretation resting on a single derivation. The schema requires it to carry a `tentative: true` flag and an explicit `caveats:` block naming how it could be wrong. *Emergent* is a claim that precipitates only at the intersection of two existing nodes, not in either alone. *Equivalency* is a bridge to an external theoretical framework — a way of saying *this pattern instantiates something already well-described elsewhere*. *Open* is for the questions you have wondered about without answering, kept first-class so they do not quietly collapse into novels.

Every node carries a provenance triple: who said it, what it rests on, how it was derived. Every edge carries one too.

Here is what that distinction looks like in YAML:

```yaml
- id: O01-first-three-months
  type: observation
  name: "First three months in new city — isolation and overwhelm"
  statement: |
    Sep–Nov 2024: working long hours, child struggling at school,
    no friends yet, stopped the running routine.
  provenance:
    attribution: { source: "Alex, self-report", date: "2024-12" }
    evidence: { type: self-report }

- id: O04-grades-recovered
  type: observation
  name: "Child's grades recovered in spring semester"
  statement: |
    Spring 2025: grades recovered to pre-move levels, coinciding with
    Alex re-establishing her own routine.
  provenance:
    attribution: { source: "Alex + report cards", date: "2025-05" }
    evidence: { type: external-record }

- id: P01-routine-as-regulation
  type: overlap
  name: "Physical routine is load-bearing for Alex's stability"
  statement: |
    When routine breaks down, other things deteriorate proportionally;
    when it returns, they stabilize. Not preference — structural.
  provenance:
    evidence:
      type: pattern-across-cases
      references: [O01-first-three-months, O04-grades-recovered]
```

```
    derivation:
      from: [O01-first-three-months, O04-grades-recovered]
      method: "induction across independent instances"

- id: N01-isolation-as-early-warning
  type: novel
  tentative: true
  name: "Isolation is an early-warning signal, not a neutral state"
  statement: |
    PROPOSED: for Alex, extended periods without meaningful social
    contact appear upstream of routine breakdown.
  provenance:
    evidence:
      type: derived-inference
      references: [O01-first-three-months]
  caveats: |
    Could be wrong: O01 conflates isolation with several other
    changes. Only one detailed episode on record; needs an
    independent second.
```

Two observations. One overlap grounded in both. One novel grounded in only the first, flagged tentative, with a caveat that names what could falsify it. The overlap could become stronger with more instances. The novel cannot become a pattern until a second, *independent* episode arrives — not the same claim restated, but a different event with the same shape.

The operating rule, adapted from Patrick McCarthy's open-knowledge-graph, is:

> ***Attribution ≠ confidence.***

This is the move that does the work.

Repetition feels like corroboration. It isn't. If Alex has said across six conversations that she "stays in misaligned situations because she is afraid of burning the relationship," that is one derivation repeated six times, not six pieces of evidence. The model agreed politely each time. Nothing new has landed.

Real confidence accumulates only from *independent* grounding: different episodes, different contexts, different evidence types. The schema forces this into the structure of the memory itself. A node whose evidence is `derived-inference` from a single episode cannot quietly become a pattern. It can only move from `novel` to `overlap` when a new, independent observation lands.

This sounds bureaucratic. It is the opposite.

Without it, a model that is polite and attentive drifts into a subtle kind of hallucination — confident about things that rest on thin inference, because those things have been said and not objected to. With it, the model can tell you the load-bearing observations — the ones most of your interpretations rest on, so a wrong episode doesn't corrupt downstream claims. It can tell you the fragile ones, with their explicit caveats. It can preserve the open questions you have been quietly answering with plausible-sounding novels.

The scaffold is open, MIT-licensed, at **github.com/parrik/know-thyself**. Paste `START_HERE.md` into a conversation with a model that has meaningful memory of you, and it walks through the construction in phases: inventory references and observations; identify patterns; name the novel interpretations tentatively; find the emergent ones; preserve the open questions; add equivalency bridges if relevant; name the practices you have adopted from all this. You get a typed YAML graph, a visual diagram, and a list of which observations a correction would cascade from.

None of the pieces are new. The provenance triple, the confidence tiers, and the emergent-at-intersection framing come from Pat McCarthy's scientific-claims schema. The *warrant* field — the reasoning between evidence and claim, stated as a separable assumption — is from Toulmin. Atomicity and link density come from Luhmann's Zettelkasten. Revisions borrow from W3C PROV-O. Effort and genre tags come from the rationalist epistemic-status convention. What the scaffold adds is a first-class *observation* node, because personal graphs, unlike scientific ones, treat episodes as things that get reinterpreted, not just as evidence for propositions.

What it produces is closer to an older thing: a Renaissance commonplace book. Structured personal notes, typed and linked, organized for retrieval and return. The difference is that a commonplace book was private. This one is designed to be readable by an AI you are talking with.

The Delphic maxim γνῶθι σεαυτόν — *know thyself* — was carved on the temple wall as advice to visitors before they consulted the oracle. The oracle is the interlocutor; know-thyself is the preparation for being understood by one.

If we are going to keep having long conversations with systems that remember us, the question of whether *we* know what they know about us, and whether they know how they know it, is not decorative.

It is the thing.