The Resonance Gate: endogenous confidence from retrieval geometry in a vector-symbolic memory

Retrieval is a membership signal, not an error detector

Jon-Paul Cacioli
Independent Researcher, Melbourne · synthium@hotmail.com
Code: github.com/synthiumjp/resonance-gate · Registration and reports: osf.io/95e2q
Preprint · July 2026

Abstract

Most systems that give a language model long-term memory or a sense of its own limits add an exogenous confidence estimator: a second model, a trained head, or an LLM judge that inspects an answer after it is produced. We ask whether confidence can instead be endogenous, a structural property of the retrieval operation itself. We build the Resonance Gate (RG), a vector-symbolic (VSA/MAP) memory substrate in which retrieving an answer and scoring confidence in it are one operation, read from the geometry of a cleanup. A small frozen transformer verbalises gate-approved content and is never consulted for facts. No confidence-specific training exists anywhere in the system.

We report a pre-registered confirmatory evaluation, a self-commissioned adversarial audit of it, and a harder fair-corpus re-test, all made public before this paper. Three results hold. An untrained closed-form mapping of the retrieval geometry to a belief/disbelief/uncertainty opinion carries essentially all the correctness information a trained logistic readout of the same features extracts (gap 0.013 AUROC2 on the clean corpus, inside a pre-registered 0.02 margin; a gradient-boosted foil does no better). Under a verify-then-speak discipline, ungrounded assertions in templated output are eliminated (0 of 263 templated bodies; 1 of 288 total, in free-text phrasing), measured by a judge validated at unit precision and recall on negation, implication, and temporal assertions. The endogenous signal is a store-membership signal, not an error detector: it separates what is in memory from what is not (gate minus a zero-parameter exact-key oracle, +0.077 AUROC2, CI excludes zero), but among items it answers it ranks correctness at chance (0.516), and membership plus one geometry scalar accounts for its whole advantage. The clean-corpus ambiguity and calibration ceilings we first measured (both 1.00) were artifacts of item construction, which the audit exposed. We give the substrate design, a load- and codebook-normalised resolution statistic with a capacity approximation, a two-geometry encoder, and the transparency process behind the corrections.

1. Introduction

Language models represent more about their own correctness than their words reveal. Correctness is decodable from intermediate activations at above-chance and often high reliability, yet this signal does not reliably reach the point of report. Our prior work localised this "know–say gap" and showed it is a routing problem: a linear probe on a model's own hidden states, projected onto the confidence-token logits with about ten weights and no fine-tuning, makes the model verbalise calibrated confidence, whereas fine-tuning the model to emit confidence leaves it near chance (Cacioli 2026a); scaling this probe-targeted approach across eight models and confirming its effect with activation patching supports the routing interpretation (Cacioli 2026b). A separate line establishes that confidence signals must be *validity-screened* before they can be interpreted at all: signals that look calibrated in aggregate can be dominated by response style rather than item-level discrimination (Cacioli 2026d). Those results point at an architectural question rather than a training one: if knowing and

saying occupy separate subsystems joined by a lossy bridge, what happens in a system with *no separate saying-subsystem to route to*, one in which the confidence simply *is* the retrieval?

This paper builds and tests such a system, and reports precisely what its endogenous confidence can and cannot do. Knowledge lives in a vector-symbolic memory substrate; a small language model is demoted to a phraser that may only verbalise what the substrate resolves. Confidence is not estimated after generation; it is read from the geometry of the retrieval. We set out to show that this arrangement yields metacognition: correctness-awareness without training. It does not. What it yields instead is narrower: a reliable signal for whether something is *in memory*, and a chance-level signal for whether a retrieved answer is *correct*. We reached that conclusion by auditing and re-testing our own initially stronger claims, and we report the negative result as the finding.

1.1 Hypotheses and how they fared

We registered five decision rules on OSF (osf.io/95e2q) after freezing the artifact and before collecting confirmatory data. We state them here in full, because a reader should not have to leave the paper to know what was tested; §5 reports each verdict, and §4 describes the firewall that separates the tuning from the test.

- **H1 (endogeneity).** The endogenous gate scalar discriminates correct from incorrect answers better than the model's own verbalised confidence (paired-bootstrap CI of the AUROC2 difference excludes zero). *Registered primary.*
- **H2 (parity with a trained readout).** An untrained closed-form mapping of the gate's retrieval features comes within 0.02 AUROC2 of a logistic head trained on those same features (development data only).
- **H2b (an architectural prediction).** A probe on the frozen mouth's logits decodes correctness at chance (CI contains 0.5), because the mouth holds no facts.
- **H3 (ambiguity typing).** The gate separates referential ambiguity (underspecified query) and stored collision (two facts on file) from clean answers, each above a 0.90 AUC floor.
- **H4 (gate discipline).** Ungrounded assertions in verified output stay at or below 2%.

The letter of all five held in the registered run. The audit (§4, §5) then showed that three of the five headline numbers were substantially determined by how the evaluation corpus was built, and a fair-corpus re-test revised their interpretation. The two that survived unqualified (H2 and H4) became the paper's positive contribution; H1 and H3's *interpretation* changed from "metacognition" to "membership," set out in §5.3.

1.2 The organising idea

We frame the design cybernetically. Ashby's law of requisite variety (Ashby 1956) and the Conant–Ashby good-regulator theorem (Conant & Ashby 1970) hold that a regulator can counter only disturbance it has the internal variety to model. A bounded substrate has bounded variety; external tools and search are *borrowed* variety; and "knowing one's limits" is the detection that incoming query variety exceeds internal variety. In RG that detection is a scalar read from the cleanup. Our results sharpen exactly which face of ignorance the substrate regulates (presence in memory) and which it does not (correctness among what is present).

1.3 Contributions

1. A memory substrate in which a confidence-bearing opinion is a closed-form function of retrieval geometry, with no confidence-specific training anywhere in the system.
2. A load- and codebook-normalised resolution statistic and an associated capacity approximation, $k_{max} \approx D/(\pi \ln N)$, derived and empirically checked, with a registered negative result on dimension scaling.
3. A two-geometry encoder (a semantic interface layer and an orthogonalising substrate layer) that resolves a measured tension between natural-language generalisation and retrieval separability.

4. The central empirical result: an untrained mapping matches a trained readout of the same features, *and* the scoped negative result that this signal detects membership rather than error, established against a zero-parameter oracle baseline on a confusable corpus under strict scoring.
5. A worked example of transparent, AI-assisted solo research: pre-registration after freeze, a public append-only notebook, two logged deviations, an accidental hash-attested replication, and a self-commissioned adversarial audit that corrected three of the author's own claims before publication.

---

2. Related work

Agent memory. A large 2024–2026 literature equips LLM agents with long-term memory: MemGPT/Letta, Mem0, Memory OS (Kang et al. 2025), HippoRAG (Jiménez Gutiérrez et al. 2024), and a dense 2026 crop including dual-process memories with an LLM quality gate (D-Mem; You et al. 2026), agentic memory with dynamic linking (A-MEM; Xu et al. 2025), structured admission control that scores candidate memories before storage (A-MAC; Zhang et al. 2026), and on-device neuro-symbolic memory operating systems (Aeon; Arslan 2026). In each, the reliability signal is *exogenous*: an LLM judge or learned scorer inspecting content after retrieval or before storage. RG differs in mechanism rather than vocabulary: its confidence is identical to the retrieval operation.

These systems share a pattern. Where they compute a "confidence" or "factual reliability" signal, that signal is almost always a measure of *grounding in the source* rather than of *truth*. A-MAC's factual-confidence factor, for instance, is the ROUGE-L overlap between a candidate memory and its supporting conversation turns, a check of whether the claim is present in and supported by the source text, not whether it is correct (Zhang et al. 2026). This is the exogenous analogue of what we find endogenously: retrieval-derived confidence, whether computed by overlap at write time or by cleanup geometry at read time, indicates *presence and support* far better than correctness. The fair-corpus test in §5.3 makes this precise for our substrate, asking whether the geometry carries correctness information beyond store membership, and finds that without training it does not.

Retrieval uncertainty and abstention. A mature literature uses retrieval signals for confidence, selective answering, and abstention, and RG must engage it directly, because a resolution-threshold gate resembles a retrieval-score threshold. Soudani et al. (2025) give an axiomatic analysis showing that standard uncertainty estimators cannot reliably assess correctness in RAG: a generator sits downstream of retrieval and can add, distort, or ignore the retrieved content, so retrieval score and answer correctness dissociate through the generator's degrees of freedom. RG's verify-then-speak discipline removes those degrees of freedom, since the mouth cannot assert beyond what resolved, which is why the honesty result (§5.2) holds. But our fair-corpus test (§5.3) surfaces the converse limitation in a new form: with the generator's freedom removed, answer correctness reduces to *resolution* correctness, and resolution correctness turns out to be largely *store membership*. The Soudani result and ours are two faces of one lesson: a retrieval score is a strong indicator of whether supporting evidence exists and a weak indicator of whether the answer is right. We situate our finding inside this literature rather than presenting it as a stand-alone property of VSAs.

The abstention lineage this connects to is worth naming explicitly, because RG's controller is an abstention mechanism: selective prediction and selective generation, conformal abstention, and learning-to-defer all show that refusing to answer under uncertainty improves reliability. Recent work moves that decision from answer-time to memory-injection time, arguing that a wrong memory silently biases every downstream step and so false-positive memory admission deserves its own metric (e.g. abstention-aware memory retrieval; the "right idea on the wrong object" observation of the coding-agent memory line). RG sits at a third point: the abstention decision is made on the *resolution* of a structured retrieval, and our contribution is a characterisation of what that decision object can and cannot support: presence reliably, correctness not without training.

Prior identifications of the problem, and a boundary from our own programme. The unreliability of memory-derived confidence has been identified from several directions. Work on memory consolidation shows that compressing a tentative claim into a stored assertion strips the framing that marked it unverified, so the system later emits a stale value confidently rather than abstaining, and a lossy memory that kept a wrong conclusion is worse than an empty one (the manufactured-confidence argument, which the authors frame explicitly as a diagnosis and its bound rather than a defense: the agent obeys the manufactured confidence regardless of source). The STALE benchmark shows frontier models and memory frameworks recognise an invalidated stored belief only about half the time, exposing a gap between *retrieving* updated evidence and *acting* on it (Chao et al. 2026). And an incentive-theoretic account traces hallucination to training that rewards confident guessing over admitting ignorance (Kalai et al. 2025). These make the *problem* case. RG contributes the *mechanism* case (why a retrieval-geometry signal cannot close the gap, because retrieval success is dominated by membership) and the *decomposition* that measures how much of the signal is membership versus correctness (§5.3). Our verify-then-speak discipline (§5.2) is, in these terms, a structural response to the incentive Kalai et al. identify: the mouth cannot be rewarded for a confident guess it is not permitted to make.

The membership-versus-correctness boundary is, moreover, a recurring finding in our own prior work rather than a one-off. The Competence Gate router (Cacioli 2026e) found that a parametric-competence signal and an evidential-grounding signal decode from different layers and do not substitute for each other ("a probe validated for one carries no usable signal for the other"), and that a competence probe applied to a grounded-QA task actively interferes. RG reproduces the same construct-specificity boundary in a third setting: the retrieval geometry that reliably signals *is this in memory* does not signal *is the retrieved answer correct*. We take the recurrence across architectures as evidence the boundary is structural, not incidental to any one design.

VSA and the VSA–LLM bridge. RG builds on classical vector-symbolic architectures (Plate 1995; Kanerva 1988, 2009; Gayler 2003; Kleyko et al. 2023a,b), subjective logic for the opinion triple (Jøsang 2016), and resonator networks for structured factoring (Frady et al. 2020). VSA-derived confidence has been used before, for example hyperdimensional uncertainty quantification in multimodal sensor fusion, but for *perceptual* uncertainty rather than memory, and to our knowledge RG is the first to make a VSA cleanup the joint substrate for memory and confidence in a language setting. The feasibility of mapping an LLM's residual stream into a VSA-structured space was recently demonstrated as an interpretability tool by the Hyperdimensional Probe (Bronzini et al. 2025), through a *trained* neural translator; we reuse the idea only as a deferred research option, and our substrate does not depend on it (§3.1).

Metacognition psychometrics. We measure the confidence signal with the type-2 signal-detection apparatus standard in human metacognition research: type-2 AUROC2 and meta-d′/M-ratio (Maniscalco & Lau 2012; Fleming & Lau 2014), with the low-sensitivity exclusions of Guggenmos (2021), applied to language models as in our own prior signal-detection work (Cacioli 2026c). To our knowledge this apparatus has been applied only to exogenous confidence signals; applying it to an architecture where the signal is the retrieval operation is part of the contribution. It also inherits a validity concern from our own prior work: a type-2 instrument measures metacognition only where the underlying type-1 task is genuinely non-trivial, and a confidence signal that looks calibrated in aggregate can fail item-level validity screening (Cacioli 2026d, the validity-scaling framework). §5.3 is, in those terms, a validity screen: restricting the analysis to items the system actually answers is what separates "the signal discriminates answerable from unanswerable" from "the signal knows which answers are right."

---

3. System

RG is a loop of six modules: an encoder, a VSA substrate, a resolution gate, a controller, a write path, and a frozen mouth. The substrate holds what is known and how strongly; the mouth is borrowed only to speak. We describe each in turn; every symbol is introduced before it is used.

3.1 Encoder: two geometries

Facts and queries arrive as text and are embedded by an off-the-shelf sentence encoder. The embedding is used in two distinct spaces, and separating them is one of the paper's design contributions. An **interface layer** resolves query terms and extracted strings against an entity/relation registry by *raw* embedding cosine. This is where natural-language generalisation lives (partial names, aliases, paraphrased relations) and where *referential* ambiguity ("which Tom?") appears as a small margin between the top two registry matches. A **substrate layer** then projects the resolved items into bipolar vector-symbolic space through a fixed, seeded random projection with shrinkage whitening (mixing weight $\lambda = 0.75$, chosen on substrate-side separability alone). Here semantics is deliberately *not* the objective; separability is. We discovered empirically (§3.5, §5) that the same embedding correlation that helps the interface generalise *hurts* the substrate's ability to keep distinct facts apart, so the two layers want opposite things from the geometry, and giving each its own space resolves the conflict. The residual-stream bridge of Bronzini et al. (2025) is a deferred alternative encoder, not a dependency.

3.2 Substrate

We use the MAP algebra: bipolar vectors in $\{-1, +1\}$, binding by elementwise multiplication (which is self-inverse, so unbinding is a single multiply), bundling by taking the sign of a sum, and permutation to mark role positions. A fact is stored as a record $R = \text{subj} \circledast \rho^1 \cdot \text{rel} \circledast \rho^2 \cdot \text{obj}$, where $\circledast$ is binding and ρ is permutation; the record is added into a running superposition (the L1 bundle) and appended to a disk-backed cleanup store of exact per-record entries (L2). Writes are O(1) and never re-read prior facts (measured flat from 10 to 50,000 stored items). A read unbinds the requested role and cleans up by cosine against the store; the sorted cosines $s_1 \geq s_2 \geq \dots$ are the raw material for both the answer and the confidence. When the L1 bundle exceeds its capacity, reads page to L2, which is exact retrieval rather than superposition.

3.3 The resolution gate

From the sorted cosines we form an absolute resolution $a = s_1$ (how strongly anything resolved) and a margin $m = s_1 - s_2$ (how cleanly the winner beat the runner-up), and map them to a subjective-logic opinion (b, d, u) with $b + d + u = 1$ (Jøsang 2016). Uncertainty u is high when nothing resolves strongly; belief b is high when one candidate wins cleanly; disbelief/conflict d is high when two candidates resolve near-equally. Because raw cosines shrink as the bundle fills, we do not threshold a directly. We gate on a **load-normalised** resolution $z(a)$, computed against both the expected hit signal $\mu_{hit}(k) = \sqrt{2/\pi k}$ and the expected null floor $\mu_{null}(N) = \sqrt{2 \ln N / D}$, where k is the current number of stored items and N the codebook size, both known exactly because the system wrote every item. Normalising against both moments makes the gate's calibration invariant to load by construction, a property no post-hoc estimator of raw cosines has.

3.4 A capacity approximation

The two moments meet where the expected hit signal sinks into the null floor, giving a closed-form estimate of usable capacity: $k_{max} \approx D/(\pi \ln N)$. We present this as a derived approximation for this regime (MAP algebra, bipolar vectors, this encoder and projection), validated empirically rather than as a universal law. At $D = 8192$, $N \approx 700$ it predicts saturation near 400, matching observation, and it yields the L1→L2 paging threshold as a derived quantity rather than a tuned constant.

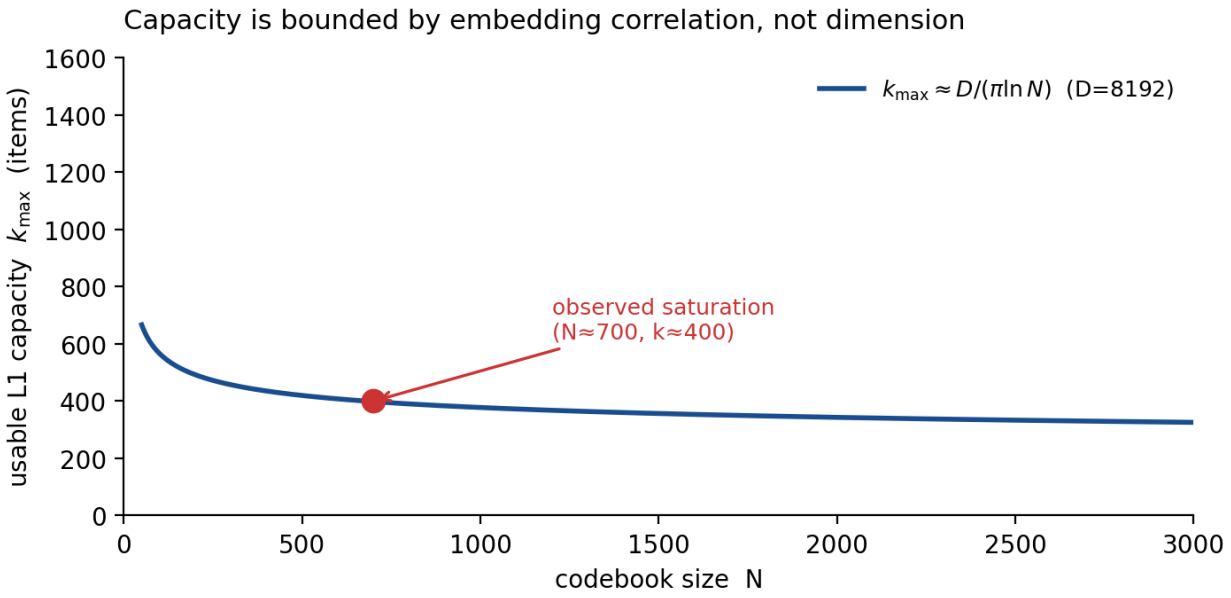

Figure 2. Usable L1 capacity k_max ≈ D/(π ln N) at D=8192, with the observed saturation point (N≈700, k≈400). Doubling D does not double capacity: the null floor is set by embedding correlation, which is dimension-invariant.

A registered negative result qualifies it: doubling D to 16384 does *not* double usable capacity, because the null floor is dominated by embedding correlation that survives sign-quantisation and is therefore dimension-invariant. Correlation, not dimension, is the binding constraint, a point worth flagging to anyone scaling a VSA memory on real embeddings.

3.5 Two-source ambiguity

RG distinguishes two kinds of "which one?". *Referential* ambiguity (the query underspecifies among registry entries) is detected as a small top-2 margin in the interface layer. *Stored* collision (two facts on file under one key) is detected as a small top-2 margin in the L2 store. On the clean corpus both separated at ceiling; §5.4 reports the important caveat that near-synonym stored collisions largely evade the key-based detector, a genuine defect of the frozen artifact rather than an evaluation artifact.

3.6 The mouth and verify-then-speak

A frozen SmolLM3-3B verbalises gate-cleared content. Factual assertions are rendered through slot-filled templates over the resolved record, which is leak-proof by construction; free generation is reserved for connective phrasing. We found that at 3B scale an ungoverned mouth fabricates in a large fraction of its *connective* text, so honesty must be structural rather than behavioural: we apply verify-then-speak to lead-in text as well, at a measured cost in fluency (§5.2, §6). The mouth is never queried for facts, and §5.1 confirms its logits carry no decodable correctness signal, as the architecture predicts.

3.7 Write path

Every record carries a bound provenance role (user-stated, assistant-inferred, or tool-derived). Version 0 writes only explicit user assertions parsed into triples, conservative by design. An echo-check reads each record back through the gate immediately after writing, so encoding failures surface when they are cheap to fix. Retraction subtracts the exact record from the bundle and tombstones the L2 entry (self-inverse binding makes the subtraction exact); an update writes a superseding record. These operations held in testing: deleted facts produce ignorance, and superseded facts resolve to their successors.

4. Method: a transparency-first evaluation

The evaluation was designed so that its negative results would be as credible as its positive ones. The mechanism is a hard firewall between exploration and confirmation, plus a public record of every departure. We describe it in enough detail to be reproduced; as with our earlier work, the process is part of the contribution.

Exploratory build, then freeze. All design and tuning happened before any confirmatory data existed, logged in an append-only notebook. At the freeze the artifact was version-tagged and every tunable recorded (dimension, whitening weight, gate parameters, templates, model hashes, projection seeds).

Pre-registration after freeze. Only then was the confirmatory study registered on OSF: the hypotheses of §1.1, their decision rules, exclusion criteria, analysis-script hashes, registered seeds, and the frozen-artifact hash. A guard prevented the analysis from running until the registration URL and timestamp were recorded.

Confirmatory run, two deviations, an accidental replication. The registered analysis was run once. Two deviations were logged and filed publicly. Deviation 1: a numerical bug in the trained-foil baseline (features constant on the development corpus produced a degenerate standardisation on the confirmatory corpus) invalidated that one comparison; it was patched instrument-side, the foil re-fit on unchanged development data, and fresh seeds drawn, with no decision rule or hypothesis changed. Deviation 2: a premature execution of the re-run on a second seed family before filing; the report was quarantined unread and hash-attested, and the procedure changed so that the human, not the agent, triggers the analysis. The quarantined run was released after the valid run and, on later reading, agreed with it on every verdict: an unplanned, hash-attested replication on an independent seed family.

Self-commissioned adversarial audit. Before writing this paper we ran an adversarial audit of the registered analysis from a clean clone, instructed to break the result. It reproduced every LLM-free number bit-exactly and confirmed the statistics with an independent implementation, but found that three of five registered headline numbers were substantially guaranteed by how the evaluation corpus had been constructed (§5). We treat the audit's findings as first-class results and corrected the public record before this preprint.

Fair-corpus re-test. The audit specified the test the confirmatory corpus had lacked. We built a harder, confusable corpus (near-synonym relation collisions, near-duplicate keys, manufactured answered-errors), adopted strict scoring, gave the verbalised baseline an actual channel to the retrieved record, and added the correct null: a zero-parameter exact-key store-membership oracle. This oracle plays the role that a scrambled-control condition plays in a factorial design: it is the baseline that decomposes "the gate is a good confidence signal" into "the gate knows what is in the store" versus "the gate knows which answers are right." Alongside it we apply a validity screen in the sense of our prior instrument-validity work (Cacioli 2026d): the metacognition question is well-posed only on items where a type-1 answer was actually forced, so we report AUROC2 both overall and restricted to answered items, and treat the answered-only quantity as the one that speaks to metacognition. This re-test (§5.3) is exploratory and labelled as such; it characterises the frozen artifact.

The full hypotheses, decision rules, deviation log, audit report, and fair-corpus code are in the repository and the OSF record.

---

5. Results

We report results in order of evidential strength after audit, not in the order registered. Registered numbers are unchanged and independently reproduced; what changed is their interpretation. Table 1 summarises that change before the details.

Table 1. What the audit and fair-corpus re-test did to each registered claim.

Registered claim	Registered verdict	After audit / fair test
H1 endogeneity (gate > verbalised)	pass (0.979 vs 0.493)	letter holds, but the verbalised baseline is architecturally unloseable; the gate's real advantage over a membership oracle is +0.077, and it is a *membership* signal (§5.3)
H2 untrained ≈ trained readout	pass (gap 0.013)	**survives**; provenance byte-verified; a boosted foil does no better (§5.1)
H2b mouth probe at chance	pass (CI contains 0.5)	holds but weak by construction; retained as design documentation, not evidence (§5.1)
H3 ambiguity typing (1.00 / 1.00)	pass	referential typing holds; stored typing was byte-identical-key construction, dropping to 0.79 on near-synonyms, 54% answered both ways (§5.4)
H4 ungrounded leak ≤ 2%	pass (0%)	**survives and strengthens** under a harder checker: 0/263 templated, 1/288 total (§5.2)

5.1 The survivor: an untrained mapping matches a trained readout

On the registered clean corpus, the endogenous gate scalar reached AUROC2 0.979 for discriminating correct from incorrect answers. The load-bearing comparison is not that number but its relation to a *trained* foil. A logistic head fit on the gate's own retrieval features (resolution, margins, k, N) using development data only reached 0.975. The untrained analytic mapping came within 0.013, inside the pre-registered 0.02 margin, and an adversarial gradient-boosted foil over the same features did no better. On this corpus the closed-form geometry-to-opinion mapping extracts essentially all the correctness information the features contain, and supervision adds nothing. The result survived the audit with the foil's provenance verified byte-for-byte.

A companion architectural prediction also held: a probe on the frozen mouth's logits decoded correctness at chance (CI contains 0.5). The mouth holds no facts, so there is nothing to decode. We flag in §6 that this operationalisation is weak (with a frozen mouth and randomly assigned facts there is no causal path to an above-chance signal, so the test cannot fail informatively), and we retain H2b as design documentation rather than evidence.

5.2 Structural honesty under a harder checker

Under verify-then-speak, ungrounded assertions in templated factual output were eliminated: 0 of 263 templated bodies leaked on the fair corpus. The single genuine leak across 288 outputs (0.35%) was in free-text connective phrasing (an event noun the lead-in filter did not ban), which is exactly the surface the design does not template. Honesty is therefore structural: the templated surface is leak-proof by construction, and the residual risk lives precisely in the free-generation surface. This result strengthened under audit. The original leak checker shared vocabulary with the emitter's own filter and missed negated, implicational, and temporal assertions, so its perfect score was partly two regexes agreeing. The fair-corpus checker uses a strong external judge, characterised at precision 1.000 and recall 1.000 on a 60-item labelled set that deliberately includes those classes. The honesty claim is thus scoped precisely: verify-then-speak eliminates ungrounded assertion on templated content, and free-text connective phrasing needs its own filter, where roughly one leak per 300 outputs remained. This places RG close to the non-hallucinating construction Kalai et al. (2025) describe in the limit: a system that answers a fixed set of questions from an exact store and otherwise outputs "I don't know." Their point is that such a system avoids hallucination precisely by declining to be calibrated-into-guessing; RG realises that construction structurally, with the store as a VSA cleanup and abstention as the gate's default when nothing resolves.

5.3 A membership signal, not an error detector

The audit and fair-corpus re-test revised the interpretation of three registered numbers. On the confusable corpus under strict scoring, strict accuracy was 0.441; of 288 items, 158 were answered, of which 59 were strict errors, fixing the small-sample problem that had limited the registered corpus (only 12 answered errors). Against the zero-parameter exact-key oracle:

- **The gate beats bare membership overall.** GATE(1−u) − ORACLE = +0.077 AUROC2, CI (+0.024, +0.128), excluding zero. The geometry carries information beyond "is the key present": chiefly, it handles unwritten-fact queries about known entities, which a pure membership test cannot score.
- **But one geometry scalar accounts for the whole advantage.** GATE(1−u) − (ORACLE + top-1 cosine) = −0.005, CI (−0.013, +0.003), containing zero. Membership plus a single resolution scalar reproduces the gate; the full subjective-logic normalisation adds nothing over it.
- **And among answered items, the gate is at chance.** Answered-only AUROC2 = 0.516; GATE − ORACLE on answered items = +0.016, CI (−0.076, +0.104), containing zero. Within the items the system chooses to answer, its confidence does not rank correctness. This restriction is the validity screen of §4: the overall AUROC2 conflates "can the signal tell answerable from unanswerable" with "can it tell right from wrong," and only the answered-only quantity isolates the metacognition claim. The overall number was high; the isolated one is chance.

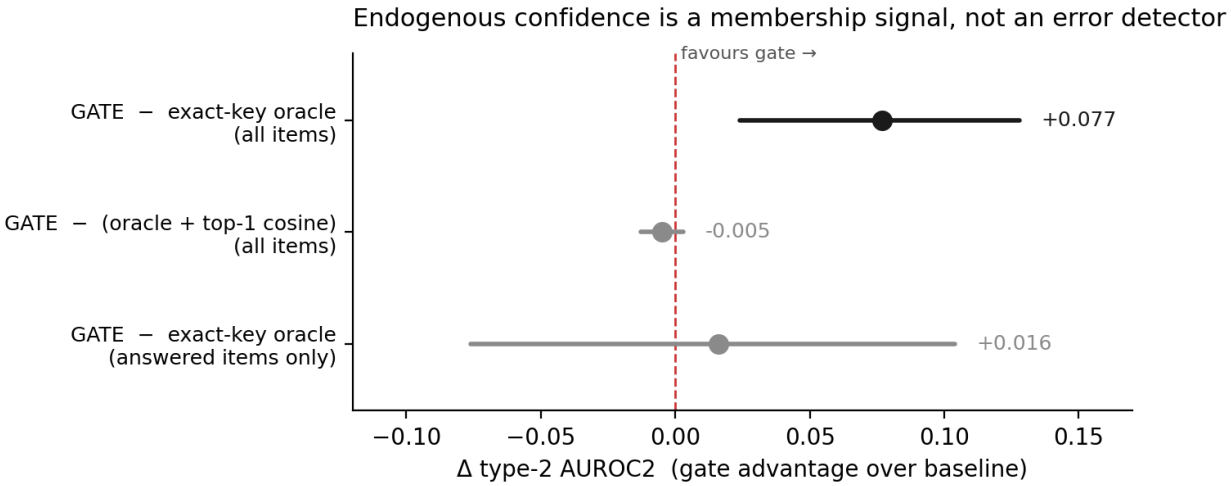

Figure 1. Gate advantage over membership baselines, type-2 AUROC2 with 95% bootstrap CIs. The gate beats a bare exact-key oracle overall (top), but ties an oracle augmented with one resolution scalar (middle) and ties the bare oracle among answered items (bottom). Dark = CI excludes zero; grey = CI contains zero.

We therefore state the finding plainly: RG's endogenous confidence is a store-membership-and-resolution signal, not an error detector. It reliably distinguishes what is in memory from what is not; it does not, without training, distinguish correct from incorrect among what it retrieves. The clean-corpus headline of 0.979 was inflated by evaluation design: the exact-key oracle alone captures 0.941 of it, and 90 of 102 incorrect items were unanswerable by construction, so most of the apparent discrimination was memory-presence rather than answer-correctness. Under strict scoring the 0.979 becomes 0.830, and the meta-d′ hypersensitivity we had registered as a predicted signature is explained by the same membership decomposition rather than by superior metacognition.

One further point sharpens the contribution and pre-empts an apparent contradiction with §5.1. The correctness signal is *present in the features but not surfaced by the untrained mapping*: on the fair corpus the trained foil reaches 0.602 answered-only where the analytic gate sits at 0.516. So §5.1 (untrained ≈ trained) and §5.3 (trained > untrained on answered items) are not in tension; they describe different corpora and different

quantities. On the clean corpus the dominant signal is membership, which the analytic mapping extracts as well as a trained one. On the fair corpus additional correctness information exists in the retrieval features, but the *analytic* mapping does not surface it. The right question is therefore not "is the geometry sufficient?" but "what can be extracted from it *without confidence-specific training?*", and the answer, for our closed-form mapping, is membership but not error. Whether some other untrained mapping could reach the answered-item signal is open (§6); training one would cross the line that defines the approach and convert RG into the exogenous-judge design it was built to contrast with.

5.4 The stored-collision defect

On the clean corpus, stored-collision detection measured AUROC 1.00. The audit showed this measured construction, not capability: planted collisions duplicated keys byte-for-byte, giving a margin of exactly zero. On fair near-synonym keys ("lives in Lisbon" versus "resides in Boston") the stored detector's AUROC falls to 0.79, and in 54% of near-synonym collision pairs the system confidently answers *both ways* with different objects. This is a real failure of the frozen artifact: same-key duplicate detection works, near-synonym contradiction detection does not. Table 2 gives representative cases. A semantic stored-margin computed in registry space, rather than by key identity, is the natural fix and is future work.

Table 2. Representative near-synonym stored collisions the frozen detector misses. *(Both facts stored; the system answers each query confidently, contradicting itself across queries.)*

Stored fact A	Stored fact B	Query	Behaviour
Maria, lives in Lisbon	Maria, resides in Boston	"Where does Maria live?"	answers one confidently; the other on re-ask
(further cases in the repository's fair-corpus report)			

5.5 Reproducibility

All LLM-free numbers reproduce bit-exactly from the registered seeds and were confirmed to four decimals by an independent implementation. Mouth-dependent baseline numbers are stable in distribution but not to the digit across GPU runs (for example 0.5005 versus a committed 0.4933), with all verdicts insensitive to deltas of that size. The frozen tree is byte-identical to its freeze tag throughout the audit and re-test. The repository ships the tunables inventory, the consumed seed families, and the BLAS-dependent caches the registered runs actually used.

---

6. Limitations

Confidence tracks memory, and within memory it tracks presence, not correctness. This is the core scoped result (§5.3), not a caveat: any application must treat the signal as "is this in the store and does it resolve," never as "is this answer right."

Synthetic corpus. All results use generated entity/relation corpora. This is appropriate for isolating the mechanism and for the controlled membership-versus-correctness contrast, but real-conversation robustness is unclaimed. The most important open question is whether the same decomposition holds on natural data with genuine paraphrase, multi-hop structure, temporal change, and contradictory sources; the stored-collision defect (§5.4) is a warning that near-synonymy alone already breaks part of the design.

The stored-collision defect is a real artifact failure, not a framing issue; near-synonym contradictions are answered both ways a majority of the time.

The mouth-probe test (H2b) is weak. With a frozen mouth and randomly assigned facts there is no causal path to an above-chance signal, and pooled logit statistics at n = 288 could not detect one regardless. We retain it as design documentation; a real test would use residual streams and deployed in-context exposure.

Template fluency ceiling. Structural honesty on templated content costs fluency; the factual register is plain and a large fraction of free-text lead-ins are rejected. A plain true sentence beats a fluent leaking one, but this bounds naturalness. We report it without a human fluency study, which is future work.

Single embedder, single 3B mouth, single corpus family, and psychometric comparability to our earlier work is approximate (the apparatus was a literature-standard reimplementation, a registered deviation). **Every conversation-level failure we observed was at the extraction/parsing boundary, not in the geometry**, which both bounds current robustness and locates where real-corpus work must focus.

---

7. Discussion

The result we set out to demonstrate, that endogenous geometry yields metacognition (error-awareness without training), did not survive a fair test. The result that did survive is more precise and more useful. Endogenous retrieval geometry is an excellent *membership* signal and a chance-level *error* signal, and a closed-form mapping extracts the membership information as well as a trained head does. The know–say gap, in this substrate, is not so much closed as relocated: by removing the mouth's authority over facts we made ungrounded assertion structurally rare (§5.2), but we did not thereby obtain correctness-awareness, because the substrate's confidence is about *retrieval*, and retrieval success is dominated by presence rather than truth.

This reframes what a VSA memory offers a larger system. It is not a drop-in metacognition module. It is a memory that reliably knows what it holds, refuses to assert what it does not, and types one kind of ambiguity. These capabilities are valuable precisely because they are structural and cheap, and an exogenous judge supplies them only probabilistically and at cost. The membership signal is a real primitive; the error signal must come from elsewhere. Placed beside Soudani et al. (2025), the picture is coherent: retrieval scores, whether read before a generator or in place of one, indicate the presence of evidence far better than the correctness of an answer.

This sits against existing work as follows. The unreliability of memory-derived confidence has been argued from several directions before us: that consolidated memories turn tentative claims into confident assertions and so should abstain but do not (the manufactured-confidence line), that models recognise invalidated beliefs only about half the time (STALE; Chao et al. 2026), and that training incentives reward guessing over "I don't know" (Kalai et al. 2025). RG contributes what these do not: a substrate in which the confidence signal is the retrieval operation itself, and a decomposition (the exact-key oracle against the gate) that measures precisely how much of that signal is membership and how much is correctness. The answer, for an untrained mapping, is that it is membership. And this is the same construct-specificity boundary we have now found in three settings: parametric competence versus evidential grounding in a tool-router (Cacioli 2026e), the same dissociation on grounded QA, and here membership versus correctness in a VSA memory. Stating that boundary cleanly, with the instrument that isolates it, is more useful than one more system reporting a high but conflated confidence number.

We also offer the process as a contribution, and here we echo a pattern from our own prior work, in which pre-registered confirmatory hypotheses were falsified and a control condition rescued the real finding. A single researcher building with an AI coding assistant can nonetheless run a study with a freeze, a public pre-registration, a logged deviation trail, an accidental replication, and a self-commissioned adversarial audit that overturned the author's own headline before publication. The audit was not a formality; it changed the paper's

conclusions. This pattern (build fast, freeze hard, register honestly, and have a hostile auditor attack the result before the field does) is the most transferable part, more than any single number.

Future work, as named sequels: a semantic stored-margin to fix the near-synonym defect (§5.4); a real-conversation corpus to test whether the membership-versus-correctness decomposition survives natural language; the residual-stream bridge (Bronzini et al. 2025) as an encoder that reads the mouth's own representations; and the open theoretical question of whether any *untrained* mapping of retrieval geometry can surface the answered-item correctness signal that §5.3 shows is present in the features but unextracted.

---

8. Reproducibility statement

Code, frozen tags (rg-freeze-1.0, rg-phase-c-1.0), the OSF registration and reports, the deviation log, the adversarial audit report, and the fair-corpus code and results are public at github.com/synthiumjp/resonance-gate and osf.io/95e2q. Model weights are third-party and referenced by pinned hash. LLM-free numbers are bit-exact from registered seeds; mouth-dependent numbers are stable in distribution. Consumed seed families are listed and never reused.

References

Arslan, M. (2026). Aeon: High-performance neuro-symbolic memory management for long-horizon LLM agents. *arXiv:2601.15311*.

Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall.

Bronzini, M., Nicolini, C., Lepri, B., Staiano, J., & Passerini, A. (2025). Hyperdimensional Probe: Decoding LLM representations via Vector Symbolic Architectures. *arXiv:2509.25045*. Code: github.com/Ipazia-AI/hyperprobe.

Chao, H., Bai, Y., Sheng, R., Li, T., & Sun, Y. (2026). STALE: Can LLM agents know when their memories are no longer valid? *arXiv:2605.06527*.

Cacioli, J.-P. (2026a). Repairing the know–say gap: a no-finetuning probe-to-logit confidence controller. *Zenodo* 10.5281/zenodo.21237443.

Cacioli, J.-P. (2026b). Making LLMs say what they know: probe-targeted fine-tuning for verbal confidence calibration. *Zenodo* 10.5281/zenodo.20436841.

Cacioli, J.-P. (2026c). Do LLMs know what they know? Measuring metacognitive efficiency with signal detection theory. *arXiv:2603.25112*.

Cacioli, J.-P. (2026d). Before you interpret the profile: validity scaling for LLM metacognitive self-report. *arXiv:2604.17707*. (See also the protocol and criterion-validation papers, arXiv:2604.17714 and arXiv: 2604.17716.)

Cacioli, J.-P. (2026e). Competence Gate: a metacognitive tool-router for Qwen3.5-4B. Model and technical release, github.com/synthiumjp / huggingface.co/synthiumjp/competence-gate-qwen3.5-4b.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2), 89–97.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443.

Frady, E. P., Kent, S. J., Olshausen, B. A., & Sommer, F. T. (2020). Resonator networks I. *Neural Computation*, 32(12), 2311–2331.

Gayler, R. W. (2003). Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience. *Proc. ICCS/ASCS Joint International Conference on Cognitive Science.*

Guggenmos, M. (2021). Measuring metacognitive performance: type-1 performance dependence and test–retest reliability. *Neuroscience of Consciousness*, 2021(1), niab040.

Jiménez Gutiérrez, B., Shu, Y., Gu, Y., Yasunaga, M., & Su, Y. (2024). HippoRAG: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37.

Jøsang, A. (2016). *Subjective Logic: A Formalism for Reasoning Under Uncertainty.* Springer.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. *arXiv: 2509.04664.*

Kang, J., Ji, M., Zhao, Z., & Bai, T. (2025). Memory OS of AI agent. *Proceedings of EMNLP 2025*, 25961–25970.

Kanerva, P. (1988). *Sparse Distributed Memory.* MIT Press.

Kanerva, P. (2009). Hyperdimensional computing: an introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1, 139–159.

Kleyko, D., Rachkovskij, D. A., Osipov, E., & Rahimi, A. (2023a). A survey on hyperdimensional computing aka vector symbolic architectures, Part I. *ACM Computing Surveys*, 55(6).

Kleyko, D., Rachkovskij, D. A., Osipov, E., & Rahimi, A. (2023b). A survey on hyperdimensional computing aka vector symbolic architectures, Part II. *ACM Computing Surveys*, 55(9).

Manufactured confidence: how memory consolidation turns hearsay into confident facts (2026). Collapse Index. *arXiv:2606.29279.* Code: github.com/collapseindex/manufactured-confidence.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422–430.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3), 623–641.

Soudani, H., Kanoulas, E., & Hasibi, F. (2025). Why uncertainty estimation methods fall short in RAG: an axiomatic analysis. *Findings of the Association for Computational Linguistics: ACL 2025*, 16596–16616.

Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., & Zhang, Y. (2025). A-MEM: Agentic memory for LLM agents. *arXiv:2502.12110.* (NeurIPS 2025.)

You, Z., Yuan, J., & Cai, J. (2026). D-Mem: A dual-process memory system for LLM agents. *arXiv: 2603.18631.*

Zhang, G., Jiang, W., Wang, X., Behr, A., Zhao, K., Friedman, J., Chu, X., & Anoun, A. (2026). Adaptive memory admission control for LLM agents (A-MAC). *ICLR 2026 Workshop MemAgent.* arXiv:2603.04549.

SmolLM3-3B (2025). Hugging Face. Open weights, data, and training code.