WRONG WITH CONVICTION: WHY CONFIDENT ERRORS EVADE DETECTION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A language model can be wrong and confident simultaneously, and the confident case matters for deployment: fluent, assertive, and invisible to signals designed to screen low-confidence output. We identify a structural reason for this blindness while showing that the model still retains internally accessible information about whether its answer is correct. We prove an input-invariance theorem for an idealized random-matrix transformer model: when input-dependent gating affects a sublinear number of dimensions and has bounded operator norm, the normalized per-layer transition-Jacobian participation ratio converges to an input-independent constant. Its endpoint-blindness corollary implies that unsupervised macroscopic geometric or spectral readers see the class-dependent part of this geometry diluted by model width below the measurement noise floor, leaving them operationally at chance, with the dilution deepening as models widen. On a nonlinear network, participation-ratio medians differ by 0.87% (AUROC 0.531). Correspondingly, semantic entropy detects genuine uncertainty (AUROC 0.774) but is near chance on adversarial confident error (0.476–0.563 across twelve models and four families). Crucially, blindness is not absence of information: our entity-grouped, out-of-fold commitment-band probe recovers a truth-related direction from frozen last-token residual states, detecting confident error at AUROC 0.731–0.849 across the powered model ladder and retaining AUROC 0.720 [0.674, 0.762] specifically on errors semantic entropy misses. We convert this direction into a Governor that identifies the error regime, routes a regime-matched correction, and refuses when residual risk remains high. Routing is necessary because the wrong regime's fix actively harms. Across five models and three families, routing beats the best single fix by +4.7 to +10.5 points, with every bootstrap interval excluding zero and McNemar $p < 0.001$ throughout; routing every item to its regime-matched corrector improves accuracy by +13.0 to +20.4 points over base. Intervention analysis reveals an asymmetric subspace where corruption survives orthogonalization while rescue collapses. Together, these results expose a separation between outward confidence, coarse internal geometry, and recoverable correctness information, showing that confident errors can be diagnosed and acted on rather than treated as terminal failures.

1 INTRODUCTION

A deployed language model occasionally produces answers that are fluent, well-formed, and internally consistent, yet wrong. Such confidently wrong answers carry real cost, especially when they occur in high-consequence domains such as a clinic, a courtroom, a flight deck, a battlefield, or a financial back office. Trustworthy use of language models therefore requires accurate characterization and mitigation of this confident-but-wrong case. We show that this failure is not the model hedging or refusing, but instead the model committing. Output-level safeguards are built for the opposite case. Confidence thresholding, sampling-based uncertainty, and abstention heuristics all assume that when a model is about to be wrong it betrays some hesitation: a flattened probability, a scattering of sampled continuations, a measurable spike in uncertainty. Yet these dangerous errors are precisely the ones that violate this assumption. The model is wrong *and* certain, and nothing at the output flags it. Filtering low-confidence generations does not remove these cases because they are not low-confidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Confident errors are not a single phenomenon; they fall into one of two categories, a *trap* or a *gap*. A trap occurs when a model errs despite possessing the relevant knowledge, because the prompt activates an absorbed misconception; that is, a common but false answer competes with and defeats the correct one. A gap occurs when the model errs because it never learned the fact at all. Recent work establishes that these two regimes, hallucination despite knowledge and hallucination from ignorance, are represented differently inside the network and call for different corrective actions (Simhi et al., 2024; 2025b): intervene on the representation when the knowledge exists; retrieve or abstain when it does not. (In our data, trap items carry a misconception-structured lure, and gap errors arise on *ordinary* items without one; Section 5.) What that work leaves open is the question a deployer must answer before building anything: (a) which regime an error belongs to, (b) whether getting it wrong is merely suboptimal or actively damaging, and (c) whether a runtime system can read the regime from internal state and route accordingly.

These observations organize a deployer's problem into three questions. First, *can a detector see a confident error at all?* Second, *if it can, does one correction serve both error regimes, or must the system route to a regime-matched fix?* Third, *can routing, correction, and refusal (the system abstaining from answering) be performed at runtime from the model's own internal state?* The first question constrains the other two, and its answer is more negative than the literature suggests: unsupervised uncertainty and geometry scores are commonly proposed as general-purpose hallucination detectors (Farquhar et al., 2024; Chen et al., 2024; Sriramanan et al., 2024), yet on confident error we find them at chance. We show that a class of detectors, unsupervised readers of the model's macroscopic endpoint geometry, as well as semantic-entropy-based sampling detectors, is constitutively blind to confident error. This blindness is not an artifact of a weak feature or a small model: we establish through an idealized model, validated on real networks, that it follows from the input-invariance of the transition Jacobian.

This property of the invariant Jacobian, established in our theorem, shuts the door on one class of detectors when faced with confident errors, yet opens it to another. A supervised probe is not bound: it can learn a specific truth direction that lives off the invariant macroscopic spectrum, in the path the model takes to its answer rather than in the answer's endpoint. This is our central result: under an idealized model, it explains why a class of unsupervised detectors fails where supervised probes succeed (Simhi et al., 2025a; Phillips et al., 2026), and locates where a runtime system must act.

Contributions. (1) An input-invariance *theorem* for an idealized random-matrix model of a transformer block, and the endpoint-blindness bound it implies for unsupervised geometric and spectral detection, validated empirically (Section 3, Appendix A). (2) Four measurements of the bound, including a semantic-entropy double dissociation (Section 4, Appendix F). (3) The supervised detector the bound leaves room for, which detects confident errors across four families, including those semantic entropy misses, survives de-confounding and an independent re-implementation, and yields a refusal gate that beats calibrated confidence (Section 6). (4) The cost of mismatched correction, which error-regime taxonomies argue for but leave unmeasured: the wrong regime's fix can actively harm, and a router that reads the regime from internal state beats the best single fix on all five models (Section 7). (5) A subspace mechanism that explains why rescue is harder than corruption (Section 8). (6) A Governor whose correction and refusal stages are each evaluated across three families (Sections 6, 9). (7) Negative results that fix the system's operating envelope (Section 10). Appendix N glosses the technical terms used throughout.

2 Related work

A. Internal-state error detection. A line of work shows that a language model's hidden states encode whether its own output is correct. Azaria & Mitchell (2023) train a classifier on hidden activations that predicts statement truthfulness more reliably than asking the model directly; Orgad et al. (2025) show that hidden states carry more error information than the output reveals; INSIDE (Chen et al., 2024) reads the eigenstructure of the sentence-embedding covariance, and LLM-Check (Sriramanan et al., 2024) reads endpoint geometry such as the covariance log-determinant. These methods establish that the signal exists. Our contribution is orthogonal and, for the unsupervised members of this family, cautionary. We show, by a theorem for an idealized model (Theorem 1) that we validate empirically, that an unsupervised detector reading a bulk statistic of the *macroscopic endpoint geometry* is bound to chance on confident error, and we then show what a *supervised* probe must read instead. The closest

work in scope detects hallucinated entities in long-form generation with linear probes across families up to 70B (Obeso et al., 2025). The settings differ in the unit of error: in long-form generation it is an entity located token by token in free text, whereas in the forced-choice and short-answer tasks we study the model commits to a single answer per item, which gives a clean confident-error label and a clean target for correction. We discuss the bridge between the two settings in Appendix M.

B. Sampling-based uncertainty and semantic entropy. Semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) clusters sampled generations by bidirectional entailment and measures the entropy over meanings, and is among the strongest black-box hallucination detectors; semantic-entropy probes (Kossen et al., 2024) approximate it from a single pass. The method is designed for genuine uncertainty, where sampling disperses. We show, both by the contraction bridge of our theorem and by direct measurement, that confident error is exactly the regime in which sampling does *not* disperse, so semantic entropy is blinded; and we exhibit a positive control on which the same pipeline works, making the failure a double dissociation rather than a pipeline defect. Concurrent work likewise finds semantic entropy unreliable when the model is confidently wrong (Ma et al., 2025; Phillips et al., 2026; Appendix D); we differ in identifying the *structural* cause, the input-invariance of the endpoint geometry, and in showing that a supervised read of the path, not a better unsupervised score, recovers the signal on the errors semantic entropy misses. This blindness bound is complementary to impossibility results on hallucination control (Karpowicz, 2025): those bound what any controller can guarantee, ours bounds what an unsupervised endpoint detector can see.

C. Error-regime taxonomy. Simhi et al. (2024) separate hallucination-despite-knowledge from ignorance and argue they are differently represented and differently treatable; HACK (Simhi et al., 2025b) adds a certainty axis (see also Simhi et al., 2025a). We operationalize the taxonomy as trap versus gap (lure-bearing trap items versus ordinary items) and contribute what the taxonomy leaves unmeasured, namely the cost of the mismatched treatment, and the router that the cost makes necessary. The statistical inevitability of some confident error under pretraining objectives is argued by Kalai et al. (2025), which contextualizes our cross-task inversion.

D. Commitment timing and basin geometry. Akarlar (2026) reports that factual and hallucinated trajectories diverge at the first generated token and that the resulting attractor is asymmetric (corruption is single-step, rescue is multi-step); concurrent work finds that larger models misfire despite knowing the answer, sharpening commitment with scale (Yeom et al., 2026); and Cherukuri & Varshney (2026) give a basin account of confidently emitted errors. We replicate the early-commitment and asymmetry phenomena and add the representational mechanism, a subspace structure that explains the asymmetry and predicts which actuator works. Appendix D relates the work to activation steering, continuous-depth dynamics, and mechanistic interpretability.

3 THE ENDPOINT-BLINDNESS THEOREM

We begin with the structural result. This section proves, for an idealized model, why an unsupervised reader of the endpoint geometry cannot separate confident errors from correct answers, and then checks the prediction on a real network.

3.1 SETUP

Let $h_l \in \mathbb{R}^d$ be the residual stream at layer l of a pre-LayerNorm transformer block, with update $h_{l+1} = h_l + f_l(h_l)$, and let $J_l(x) = \partial h_{l+1}/\partial h_l = I + \partial f_l/\partial h_l$ be the transition Jacobian evaluated at the last-token state for input x. Write $M := J_l^\top J_l$ and let $\{\sigma_i\}$ be the singular values of J_l. We study the *participation ratio*

$$\mathrm{PR}(J_l) \;=\; \frac{(\operatorname{Tr} M)^2}{\operatorname{Tr}(M^2)} \;=\; \frac{(\sum_i \sigma_i^2)^2}{\sum_i \sigma_i^4} \;=\; d\,\frac{m_1^2}{m_2},$$

where $m_k := \frac{1}{d}\operatorname{Tr}(M^k)$ denotes the k-th normalized spectral moment of M. This quantity lies in $[1, d]$ and measures the effective number of directions in which J_l acts. The participation ratio is a natural target for an unsupervised endpoint detector because it is a single, scale-free summary of the local geometry of the transition map, exactly the kind of quantity the unsupervised spectral readers compute, namely the covariance log-determinant, EigenScore, and the participation ratio itself.

3.2 AN IDEALIZED MODEL AND ITS INVARIANCE

We decompose the Jacobian into an input-independent weight operator and an input-dependent gating term and make three modeling assumptions that abstract a transformer block into a tractable random-matrix object.

Assumption 1 (idealized block) $J_l(x) = I + W_{\text{eff}} + \Delta(x)$, where (i) W_{eff} is drawn from a rotationally-invariant ensemble with bounded fourth moments, so the empirical spectral distribution of $M^0 := (I + W_{\text{eff}})^\top (I + W_{\text{eff}})$ converges almost surely to a deterministic law μ_l and $\|M^0\|_{\text{op}} = O(1)$; and (ii) the perturbation $E(x)$ that $\Delta(x)$ induces in M satisfies $\operatorname{rank} E(x) = o(d)$ *and* $\|E(x)\|_{\text{op}} \le C$ uniformly in x.

Theorem 1 (input-invariance, idealized model) Under Assumption 1, as $d \to \infty$,

$$\text{PR}(J_l(x))/d \overset{\text{a.s.}}{\to} c_l = \frac{(\int t \, d\mu_l(t))^2}{\int t^2 \, d\mu_l(t)},$$

a constant determined by the layer's weight ensemble and independent of the input x.

Proof sketch (full proof in Appendix A). Input dependence enters only through $E(x)$, whose rank is $o(d)$ and whose norm is bounded, so by the trace inequality $|\operatorname{Tr} A| \le \operatorname{rank}(A)\,\|A\|$ it moves each normalized moment m_1, m_2 by $o(1)$; the ratio $\text{PR}/d = m_1^2/m_2$ therefore converges to the input-independent constant c_l. □

Corollary 1 (endpoint blindness) Let $s(x) = \Phi(\text{ macroscopic endpoint geometry of } J_l(x)\,)$ be an *unsupervised* score obtained from a fixed Lipschitz functional Φ of the normalized participation ratio. Under Theorem 1 the argument of Φ has vanishing input-dependent variation. If the observed score adds independent, class-independent continuous measurement noise of nonvanishing scale in these normalized units, then s has vanishing class separation relative to that noise and its asymptotic AUROC tends to 0.5.

The scope of the corollary is deliberately narrow. It bounds unsupervised readers of this normalized bulk statistic under the stated noise condition; it does not bound arbitrary spectral functionals or finite-width rank statistics. A *supervised* probe trained on h_l is explicitly *not* covered. It may read sub-macroscopic, off-spectrum directions about which the theorem says nothing. This is the central design implication, not a loophole. It motivates testing whether a supervised path-state probe can recover information that this particular bulk statistic loses, as we do in Section 6.

3.3 EMPIRICAL VALIDATION

Real transformers violate Assumption 1: input-dependent attention can, in principle, shift the macroscopic spectral density between input classes. Therefore the invariance is a property we must check rather than assume. We compute $\text{PR}(J_l)$ at the commitment band of Qwen2.5-1.5B for $n = 234$ confident-correct and 234 confident-wrong items matched on confidence. The median participation ratio is 400.3 for confident-correct and 396.8 for confident-wrong, a difference of 0.87%, and the resulting separation is $\text{AUROC} = 0.531$, an *observed*, near-chance value on a real network, consistent with the *idealized* asymptotic prediction of Corollary 1 ($\text{AUROC} \to 0.5$); the small residual margin above 0.5 is consistent with a finite-width, finite-sample effect; it does not establish a universal impossibility result (the 0.5 is a limit; 0.531 is a measurement). This does not show that no geometric reader can distinguish the classes: the theorem applies to the idealized model, while the real-network result tests one bulk statistic. Appendix B gives a conditional contractive-dynamics bridge for sampling-based detectors such as semantic entropy.

4 MEASUREMENTS OF THE BOUND

The theorem predicts that unsupervised readers are blind to confident error. We now test that prediction on the strongest sampling-based detector, semantic entropy; Appendix F adds three further unsupervised readouts. We compute Farquhar semantic entropy (Farquhar et al., 2024)

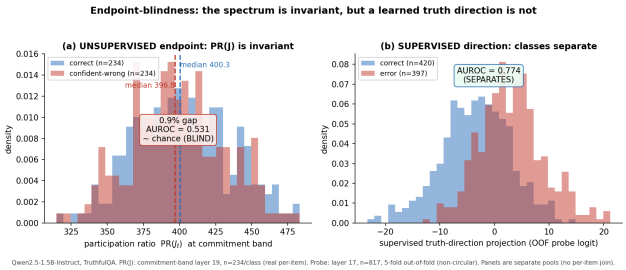

Figure 1: Endpoint-blindness on Qwen2.5-1.5B-Instruct (TruthfulQA). (a) At the commitment band the transition-Jacobian participation ratio $\mathrm{PR}(J_l)$ is invariant between classes (0.87% median gap, $\mathrm{AUROC} = 0.531 \approx$ chance; $n = 234$/class), so an unsupervised detector built on this endpoint statistic is at chance, as Corollary 1 predicts. (b) A supervised truth-direction projection (out-of-fold probe logit, 5-fold) escapes the invariance and separates error from correct ($\mathrm{AUROC} = 0.774$, $n = 234$/class; numerically coincident with, but a distinct measurement from, the semantic-entropy positive control of Section 4). Panels use separate item pools of the same model (not a paired per-item comparison). Single-model validation.

Table 1: Semantic entropy on adversarial confident error (TruthfulQA; $n = 800$ for the 7B-and-larger rows and Llama-3.2-1B, $n = 400$ for the remaining small models) against the genuine-uncertainty positive control (TriviaQA). The method is operationally useless on the former (every AUROC point estimate below 0.60) and works on the latter. AUROC is SE→error (aggregate); CI is bootstrap 95%.

Model	AUROC	95% CI	Model	AUROC	95% CI
Qwen-0.5B	0.476	[.424,.525]	Llama-3.2-1B	0.521	[.476,.565]
Qwen-1.5B	0.493	[.435,.553]	Llama-3.2-3B	0.541	[.479,.601]
Qwen-3B	0.508	[.447,.568]	Llama-3.1-8B	0.548	[.507,.588]
Qwen-7B	0.557	[.518,.599]	Mistral-24B	0.560	[.520,.599]
Qwen-14B	0.563	[.525,.604]	Gemma-2-2B	0.561	[.503,.616]
Qwen-32B	0.552	[.512,.591]	Gemma-2-27B	0.527	[.489,.565]

with bidirectional-entailment NLI clustering (`roberta-large-mnli`, $k = 8$ samples). On TruthfulQA, an adversarial misconception set in which the errors are confident, semantic entropy does not separate wrong from correct. Across twelve models and four families the AUROC ranges from 0.476 to 0.563, and every 95% confidence interval has an upper bound below 0.65 (Table 1; per-model bars in Figure 3, Appendix F.2).

Positive control. On TriviaQA, where uncertainty is genuine, the same pipeline reaches AUROC 0.774 (Qwen-1.5B, $n = 150$), and Gemma-2-27B reaches 0.687 against 0.527 on confident error, so the double dissociation holds at the largest scale. On a confident-wrong-specific split (high forced-choice margin), semantic entropy stays at AUROC $\in [0.487, 0.577]$, all < 0.61.

A non-circular split, labeled by an independent forced-choice margin rather than by the sampling signal, gives the same result (AUROC 0.468–0.481 on three Qwen models; Appendix F.2).

5 Experimental Setup

Before building the detector and the system around it, we fix the experimental ground on which every later number stands. *Data.* The detector is evaluated on a six-dataset pool of 7,710 forced-choice / short-answer items spanning MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), OpenBookQA (Mihaylov et al., 2018), and CommonsenseQA (Talmor et al., 2019); the semantic-entropy measurements use a TruthfulQA-derived set treated as an indivisible unit, with TriviaQA (Joshi et al., 2017) as the genuine-uncertainty positive control; the refusal analysis uses an earlier 3,000-item pool from the same sources.

Where we read. For every internal-state measurement we read the last-token residual stream in the *commitment band*, approximately 0.68–0.86 of relative depth, where the answer commits. The

Table 2: Per-model confident-error detection AUROC across the four-family ladder. The powered set (Qwen ladder + Llama-8B + Gemma-27B) is POWERED: the pre-registered decision rule was met, with real n, tight CIs, and cleared confound floors. Mistral-Small-24B† is reported *separately* as a marginal cross-family probe: with only $n = 85$ confident-wrong items it falls short of the pre-registered power bar, and its CI does not clear the confound floor, so it is *not* part of the powered result. The validations that follow are POWERED.

Model	AUROC	95% CI	Model	AUROC	95% CI
Qwen2.5-0.5B	0.731	[0.715, 0.746]	Qwen2.5-32B	0.821	[0.771, 0.865]
Qwen2.5-1.5B	0.779	[0.760, 0.797]	Llama-3.1-8B	0.849	[0.830, 0.866]
Qwen2.5-3B	0.762	[0.738, 0.787]	Gemma-2-27B	0.775	[0.733, 0.814]
Qwen2.5-14B	0.802	[0.760, 0.843]	Mistral-Small-24B†	0.727	[0.660, 0.787]

detector reads fixed relative depths 0.68, 0.79 and 0.86, chosen once on Qwen-1.5B from 817 TruthfulQA items (about 10.6% of the evaluation pool) and reused unchanged for every model; no layer is chosen during evaluation. No base-model weights are modified for detection, routing, or steering, and parametric correctors are adapters (Section 9).

Labels and baselines. An item is a *confident error* when the model's answer is wrong and its calibrated (temperature-scaled) confidence exceeds a threshold τ. *Trap* items are misconception-structured lures in which a common wrong answer competes with the correct one; *ordinary* items lack a structured lure; these item types are our operational labels for the trap and gap regimes of Section 1. Baselines are calibrated confidence, a random gate, and semantic entropy, the last validated as a comparator by its 0.774 positive control.

Power and labeling convention. Every quantitative claim carries an evidence label. A POWERED claim states its n and a confidence interval and meets a pre-registered power bar; an EXPLORATORY claim is small-n or lead-grade and is never presented as a finding. De-confounding is a pre-run gate. A result computed on a confounded metric is reported as confounded regardless of n.

Models. We use the Qwen2.5 instruction-tuned ladder (0.5B, 1.5B, 3B, 7B, 14B, and 32B), the Llama-3 line (Llama-3.2-1B, Llama-3.2-3B, and Llama-3.1-8B), Gemma-2 (2B and 27B), and Mistral-Small-24B (twelve models across four families), in BF16 with deterministic kernels and a fixed random seed (2026), unquantized, through HuggingFace Transformers. Not every model appears in every experiment; the per-experiment roster is reported with each result and summarized in Appendix C, and the sparse-autoencoder analysis of Appendix F.4 additionally uses Gemma-2-9B. The ladder lets us separate scale-dependent effects from architecture-dependent ones, and the four families test vendor portability.

6 THE SIGNAL THE BOUND LEAVES: A SUPERVISED DETECTOR

The bound forecloses unsupervised endpoint detection but explicitly permits a supervised probe. We train a linear probe on the commitment-band last-token residual to flag confidently-wrong risk, using entity-grouped folds so that the same entity never appears in both train and test. The number of confident errors available for training varies by model, from 85 to $1{,}715$, because the more accurate larger models make fewer of them. Table 2 reports the result; the paragraphs that follow check that the signal is reproducible, is not a confound, and improves the decision a deployer makes.

Independent re-implementation. A second implementation sharing no code with the first reaches grouped out-of-fold AUROC 0.783 at a pre-set band layer (0.787 at its best layer; LDA 0.697), a grouped-cluster bootstrap median of 0.768 with 95% CI $[0.702, 0.828]$ excluding chance, fold stability of 0.788 ± 0.009 over forty splits, and a permutation null of mean 0.500 with $p = 0.002$. The signal is reproducible across implementations, not an artifact of one training pipeline.

De-confound battery. We test the four most plausible confounds, and the probe passes all four: response length is not the driver (full-feature 0.783 against 0.591 norm-only), entity leakage is small (held-out gap $+0.051$), prevalence is irrelevant (balanced AUROC 0.785), and column-shuffling the features collapses AUROC to chance (0.502; Table 4, Appendix H).

The refusal payoff. Detection is only useful if it improves the decision a deployer actually makes: whether to answer or abstain. We cast this as a selective-prediction problem (Geifman & El-Yaniv, 2017) and compare a risk gate built on the probe against calibrated confidence, a random gate, and semantic entropy at matched coverage, restricted to the *danger zone*: items whose calibrated confidence exceeds the threshold τ, where a confidence-based gate would answer and every error is a confident one. The probe gate beats calibrated confidence by $+5.7$ to $+6.4$ points in the danger zone at matched coverage; on the full item stream the gate reaches AUROC ≈ 0.90, and a representative operating point is roughly 93% precision at 50% recall with about 4% false alarms. Correcting before refusing helps. A steer-then-refuse policy improves the selective-AUC from 0.668 (refuse-only) to 0.700. Abstention is measurably better made from internal state than from the model's own confidence.

The probe sees exactly what semantic entropy is blind to. The sharpest test of the supervised/unsupervised dissociation is to run the probe on precisely the items semantic entropy misses. On the *SE-blind* confident-wrong items (those where semantic entropy itself scores ≈ 0.25, inverted), the supervised commitment-band probe still separates error, at pooled AUROC 0.720 $[0.674, 0.762]$ on $n = 188$ with permutation $p < 0.001$, replicated across three families (Qwen-1.5B 0.710, Gemma-2-2B 0.728, Llama-3.2-3B 0.703). The probe's AUROC on this SE-blind subset stays within 0.04 of its full-pool value in this replication (Qwen 0.710 vs. 0.677, Gemma 0.728 vs. 0.744, Llama 0.703 vs. 0.727; these per-family replication runs are separate from the main detector ladder of Table 2), so conditioning on semantic-entropy blindness does not inflate it.

Supervised versus unsupervised, side by side. Every unsupervised reader we tested sits near chance on confident error: the participation ratio at 0.531 (Section 3), semantic entropy at 0.476–0.563 (Table 1), hidden-state dispersion at 0.41–0.50, and a sparse-autoencoder readout at 0.48–0.59 (Appendix F). On the same model families the supervised probe reaches 0.731–0.849 (Table 2). The information is present; only the unsupervised readers miss it.

7 Errors come in regimes: anti-transfer and the router

Detection tells a deployer that an answer is likely wrong, not how to fix it, so we now turn to correction. A *corrector* is an intervention that moves a flagged item's answer toward the correct one: a steering vector added to the commitment-band state at inference time, or a rank-8 low-rank adapter per regime (Section 9). Each is tuned for one regime of Section 1, the trap (knowledge present, lure wins) or the gap (knowledge missing); can one corrector serve both?

The wrong fix is harmful. We measure what happens when a corrector tuned for one regime is applied to the other. The scoped steering anti-transfer is -0.20 to -0.40 at scales up to 1.5B and orthogonalizes by 3B, and the transfer is directional. Steering an ordinary item into the trap direction helps ($+0.28$), while steering a trapped item out of it harms (-0.40), an easy-in/hard-out asymmetry that the mechanism of Section 8 will explain. The mismatched fix is damaging, not just suboptimal, which converts the two-regime taxonomy from a description into a routing requirement.

The router. We build the router the requirement implies. It reads the regime from the commitment-band state and applies the matching corrector. On a mixed-regime test of $n = 1{,}013$ items per model (500 trap, 513 ordinary), the router beats the best single fix across five models and three families by $+4.7$ to $+10.5$ points (Table 3), with McNemar $p < 0.001$ and every bootstrap CI excluding zero, and $+13.0$ to $+20.4$ over base. The ordering router $>$ best single fix $>$ base holds on all five models.

Table 3: The router beats the best single fix across five models and three families ($n = 1{,}013$/model; 500 trap, 513 ordinary). Every bootstrap CI excludes zero; McNemar $p < 0.001$ throughout.

Model	Family	Gain over best single fix	$\approx$ gain over base
Qwen2.5-0.5B	Qwen	$+10.5$	—
Qwen2.5-1.5B	Qwen	$+4.7$	$+13$
Qwen2.5-3B	Qwen	$+6.2$	—
Gemma-2-2B	Gemma	$+7.1$	$+20.4$
Llama-3.2-3B	Llama	$+5.3$	$+14.3$

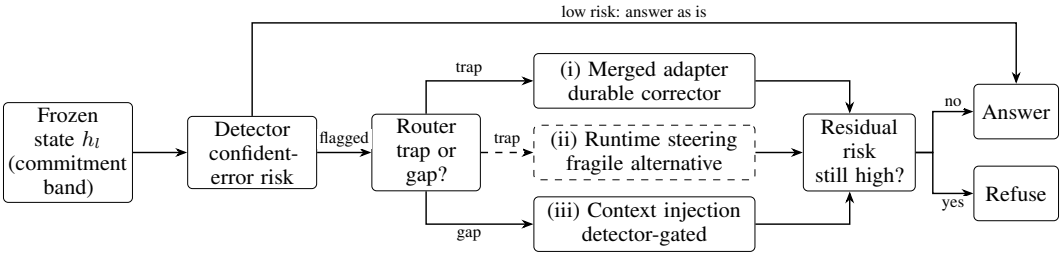

Figure 2: The Governor on frozen internal state. The detector reads confident-error risk; low-risk answers pass through. The router sends trap errors (knowledge present) to the merged adapter (i) or, less reliably, runtime steering (ii, dashed), and gap errors (knowledge missing) to detector-gated context injection (iii); if residual risk stays high, the system refuses. Section 9 reports the correction path on every item; refusal is evaluated in Section 6.

The regime representation is capacity-emergent. The re-run is leakage-controlled by leave-one-source-out routing together with a source-identity-accuracy bound, and the control reveals a scaling law. Held-out OpenBookQA routed-ordinary accuracy rises from 0.012 at 0.5B to 0.21 at 1.5B to 0.43 at 3B. The dataset-identity leakage that contaminates a naive separability claim scales *inversely* with model size, and the genuine regime representation emerges with capacity. We therefore report perfect separability as a confounded artifact, not a finding (Section 10), and rest the routing claim on the leakage-controlled router and the capacity-emergent representation.

8 Why rescue is harder than corruption: subspace structure

Routing and refusal are not an engineering convenience; they are forced by the representational geometry, and the same geometry explains the directional asymmetry of Section 7. We intervene bidirectionally along the soft gold↔lure axis ($n = 100$ right-steer and 76 wrong-steer, $\alpha = 0.6$, layer 20, with bootstrap CIs), and we repeat the intervention after orthogonalizing away the top five high-variance modes, against an isotropic control and an in-subspace control. The gold↔lure axis is strongly anisotropic (a $219\times$ mode) and separates right from wrong out of fold at accuracy 0.70.

The two directions behave asymmetrically, and the asymmetry replicates across families. The asymmetry itself is not new: Akarlar (2026) reports that hallucinated trajectories are entered in a single step but left only over several. What we add is where it lives in the representation. *Corruption survives orthogonalization*: with the dominant variance modes removed, corrupting a correct answer still works, at AUROC 0.92 against 0.35 for the isotropic control on Qwen-1.5B, 0.73 against 0.37 on Gemma-2B, and 0.96 against 0.08 on Llama-3.2-3B, three families out of three with confidence-interval separation. Corruption is therefore carried by the specific gold↔lure direction. *Rescue collapses*: pulling a wrong answer right falls to near zero in every condition on all three families. (A finer in-subspace sub-claim is Qwen-only and exploratory; Appendix I.) The reading is intuitive. Knocking a correct answer wrong needs only the specific direction, whereas pulling a wrong answer right needs the dominant subspace as well. This is the representational reason rescue is harder than corruption, and it predicts the operational facts. One-shot steers break more reliably than they repair, which argues for merged adapters over runtime steering as the durable actuator.

9 The barrier scales, and the integrated Governor

The basin persists. It remains reachable by activation patching from 1.5B to 32B, and its entry barrier steepens with scale rather than the basin vanishing (Appendix J). Because the failure does not fade as models grow, a runtime layer that handles it is needed at every scale; we now assemble it.

The Governor. The system reads the regime from the commitment-band state, applies a regime-matched corrector (a rank-8 low-rank adapter (Hu et al., 2022) per regime), and refuses when residual risk is high, all on frozen internal state with only the commitment-layer index re-tuned per family (Figure 2); a trap adapter trained category-held-out and merged into the weights recovers $+29.1$

points on unseen categories. We evaluate correction and refusal separately. Routing each item to its regime-matched adapter (two adapters trained in-run) raises forced-choice accuracy over base from 0.608 to 0.738 ($+13.0$, 95% CI $[+10.3, +15.9]$) on Qwen-1.5B, from 0.597 to 0.802 ($+20.4$, $[+17.5, +23.5]$) on Gemma-2B, and from 0.687 to 0.830 ($+14.3$, $[+11.7, +17.0]$) on Llama-3.2-3B, in one powered run of $n = 1{,}013$ items per model with every item answered; these figures include no refusal, which Section 6 evaluates as selective prediction. Even the simplest confidence signal changes what a small model can be used for: answering only its most confident half (by letter-logit margin), a small model matches or passes the next model up in its own family, 13–47$\times$ larger and answering every item, on MMLU and CommonsenseQA in all three families (Appendix J, Table 6).

Three correctors, matched to the failure. The corrective stage is not one mechanism but a small menu chosen by *why* the model is about to be wrong, and the routing decision is therefore also a corrector-selection decision. *(i) Merged adapter: the durable corrector.* For the trap regime, where the model holds the right knowledge but commits to the wrong answer, the merged adapter above repairs the commitment; once merged it is permanent and free at inference. *(ii) Runtime steering: a bounded alternative.* The same trap repair can be applied at inference time by nudging the commitment-band representation along the gold↔lure axis ($+0.233$ at $\leq$3B), but it is non-monotone in scale and the *wrong*-regime steer actively harms (-0.20 to -0.40 anti-transfer), so we report it as a fragile alternative and prefer the merged adapter. *(iii) Context injection: a different lever for a different failure.* When the model errs because it *lacks* the knowledge (the gap regime) rather than mis-commits, representation surgery cannot help; supplying evidence in the prompt recovers 29% of the small model's closed-book errors (evidence located 97% of the time) while retaining 71% of what it already had right. Because indiscriminate injection would disturb the latter, it is detector-gated.

10 Limitations

We report the system's limits as results: each fixes a component of the operating envelope and tells a deployer what to rely on. *Steering is non-monotone in scale.* The commitment-layer steer corrects $+0.233$ in the right regime at $\leq$3B and is effective only at the commitment layer, but its net effect across regimes is $+0.158$ $[0.067, 0.258]$ at 1.5B, -0.045 $[-0.075, -0.02]$ at 3B, and $+0.179$ $[0.065, 0.311]$ at 14B. This is a robust mid-scale failure with recovery at the large end, which refutes a simple attenuation story; at scale the pipeline therefore relies on adapters and refusal. *Format.* All experiments, for the detector as well as the Governor, use forced-choice and short-answer formats; the detector is evaluated on models up to 32B and the Governor's routed correction on models of 1.5B to 3B parameters. Open-generation deployment is untested here and is the study's principal limitation. *Dispersion is a non-separation, not a null.* The hidden-state dispersion result is underpowered ($n = 50$/class) to assert the absence of a signal. *Separability is confounded.* Perfect trap-versus-gap separability is dataset-identity leakage at small scale and is reported as a limitation, not a finding; the capacity-emergent routing is what survives. *Transfer is moderate.* Pooled across benchmarks the detector transfers at AUROC 0.713 (range 0.56–0.75), short of the ≈ 0.8 bar that would warrant calling it task-general. *Substrate claims are exploratory.* The dynamical-substrate observables are $N = 1$ and exploratory, and the deployable signal does not depend on them. *The theorem is for an idealized model.* Theorem 1 holds under Assumption 1; real transformers can violate it, which is why we validate the conclusion empirically rather than asserting it of arbitrary networks.

11 Conclusion

Confident errors are internally visible, early-committed, and divided into regimes, but the unsupervised tools that should catch them are blind for a structural reason. The endpoint geometry they read is input-invariant (proven for an idealized model and validated empirically). The discriminating signal lives off that geometry, in a supervised read of the path; the errors come in regimes a single fix cannot serve, and the wrong fix actively harms; the mechanism is an asymmetric subspace in which corruption sticks and rescue collapses; and the barrier scales. What this implies, and what we build and validate part by part across families, is a refusal-first, representation-read governance layer that detects the regime, routes the matching correction, and refuses on residual risk. For confident errors, one fix does not fit all, and the endpoint cannot tell you so.

Reproducibility statement

All runs use a fixed random seed (2026), BF16, deterministic kernels, and unquantized HuggingFace Transformers. The full proof of Theorem 1 with all constants is in Appendix A, the contraction bridge in Appendix B, and per-experiment configuration in Appendix C. The independent re-implementation of Section 6 used no shared code. Code and data to reproduce the reported results will be released publicly upon publication (withheld here for double-blind review), together with an interactive result audit in which every claim is recomputable in the browser and every reference is link-checked. The models are off-the-shelf open weights; we cannot exclude that the public benchmarks used here (TruthfulQA, TriviaQA) appeared in their pretraining, and we treat the TruthfulQA lures as high-probability under pretraining statistics accordingly.

Ethics statement

A refusal-first governor reduces confident misinformation from self-hosted models in air-gapped, regulated, or sovereign settings where cloud-side output filters are unavailable. The risks are an over-trusted imperfect gate, for which we report false-alarm and miss rates at operating points; misrouting of out-of-distribution inputs, for which we recommend the refuse path on router-uncertain items; and the dual-use nature of steering, where the fact that corruption is easier than rescue (Section 8) argues for weight-merged adapters over runtime steering in deployment.

AI use statement

As principal investigator, the author built and directed a multi-agent AI laboratory (Claude Code, GPT, Gemini) that assisted with proofs, experiments, analysis, interpretation, and drafting. The theory is the author's, who verified every finding against raw results and takes responsibility.

Competing interests

The author has a pending U.S. patent application related to the methods described in this paper. The author declares no other competing interests.

References

Akarlar, G. A. Hallucination as Trajectory Commitment: Causal Evidence for Asymmetric Attractor Dynamics in Transformer Generation. arXiv:2604.15400, 2026.

Azaria, A. & Mitchell, T. The Internal State of an LLM Knows When It's Lying. *Findings of EMNLP*, 2023. arXiv:2304.13734.

Bai, Z. & Silverstein, J.W. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.

Bricken, T. et al. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*, 2023.

Chen, C. et al. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. *ICLR*, 2024. arXiv:2402.03744.

Chen, J.Y.C., Lakhmani, S.G., Stowers, K., Selkowitz, A.R., Wright, J.L., & Barnes, M. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science* 19(3):259–282, 2018. doi:10.1080/1463922X.2017.1315750.

Cherukuri, K. & Varshney, L.R. Hallucination Basins: A Dynamic Framework for Understanding and Controlling LLM Hallucinations. arXiv:2604.04743, 2026.

Clark, P. et al. Think you have Solved Question Answering? Try ARC. arXiv:1803.05457, 2018.

Farquhar, S. et al. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 2024.

Fernando, T. & Guitchounts, G. Dynamics of the Transformer Residual Stream. arXiv:2605.14258, 2026.

Geifman, Y. & El-Yaniv, R. Selective Classification for Deep Neural Networks. *NeurIPS*, 2017. arXiv:1705.08500.

Guo, C. et al. On Calibration of Modern Neural Networks. *ICML*, 2017. arXiv:1706.04599.

Hendrycks, D. et al. Measuring Massive Multitask Language Understanding (MMLU). *ICLR*, 2021. arXiv:2009.03300.

Hu, E.J. et al. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 2022. arXiv:2106.09685.

Joshi, M. et al. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *ACL*, 2017. arXiv:1705.03551.

Kalai, A.T. et al. Why Language Models Hallucinate. arXiv:2509.04664, 2025.

Karpowicz, M.P. On the Fundamental Impossibility of Hallucination Control in Large Language Models. arXiv:2506.06382, 2025.

Kossen, J. et al. Semantic Entropy Probes. arXiv:2406.15927, 2024.

Kuhn, L. et al. Semantic Uncertainty. *ICLR*, 2023. arXiv:2302.09664.

Li, K. et al. Inference-Time Intervention. *NeurIPS*, 2023. arXiv:2306.03341.

Lieberum, T. et al. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. arXiv:2408.05147, 2024.

Lin, S. et al. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *ACL*, 2022. arXiv:2109.07958.

Lindsey, J. et al. On the Biology of a Large Language Model. *Transformer Circuits Thread*, 2025.

Lu, Y. et al. Beyond Finite Layer Neural Networks. *ICML*, 2018.

Ma, H. et al. Semantic Energy: Detecting LLM Hallucination Beyond Entropy. arXiv:2508.14496, 2025.

Marčenko, V.A. & Pastur, L.A. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sbornik*, 1967.

Marín, J. A Geometric Taxonomy of Hallucinations in LLMs. arXiv:2602.13224, 2026.

McDermott, P., Dominguez, C., Kasdaglis, N., Ryan, M., Trahan, I., & Nelson, A. Human-Machine Teaming Systems Engineering Guide. *The MITRE Corporation*, Bedford, MA, 2018. MITRE Case 17-4208.

Mihaylov, T. et al. Can a Suit of Armor Conduct Electricity? (OpenBookQA). *EMNLP*, 2018. arXiv:1809.02789.

Obeso, O., Arditi, A., Ferrando, J., Freeman, J., Holmes, C. & Nanda, N. Real-Time Detection of Hallucinated Entities in Long-Form Generation. arXiv:2509.03531, 2025.

Orgad, H. et al. LLMs Know More Than They Show. *ICLR*, 2025. arXiv:2410.02707.

Phillips, E. et al. Entropy Alone is Insufficient for Safe Selective Prediction in LLMs. arXiv:2603.21172, 2026.

Rimsky, N. et al. Steering Llama 2 via Contrastive Activation Addition. *ACL*, 2024. arXiv:2312.06681.

Simhi, A. et al. Distinguishing Ignorance from Error in LLM Hallucinations. arXiv:2410.22071, 2024.

Simhi, A. et al. Trust Me, I'm Wrong: LLMs Hallucinate with Certainty Despite Knowing the Answer. *Findings of EMNLP*, 2025a. arXiv:2502.12964.

Simhi, A. et al. HACK: Hallucinations Along Certainty and Knowledge Axes. arXiv:2510.24222, 2025b.

Sriramanan, G. et al. LLM-Check. *NeurIPS*, 2024.

Stowers, K., Kasdaglis, N., Newton, O., Lakhmani, S., Wohleber, R., & Chen, J. Intelligent agent transparency: the design and evaluation of an interface to facilitate human and intelligent agent collaboration. *Proc. Human Factors and Ergonomics Society Annual Meeting* 60(1):1706–1710, 2016.

Stowers, K., Kasdaglis, N., Rupp, M.A., Newton, O.B., Chen, J.Y.C., & Barnes, M.J. The impact of agent transparency on human performance. *IEEE Transactions on Human-Machine Systems* 50(3):245–253, 2020.

Talmor, A. et al. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. *NAACL*, 2019. arXiv:1811.00937.

Templeton, A. et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Yeom, J., Sok, J., Kim, H., Park, S., Park, J. & Kim, T. Hallucination as Commitment Failure: Larger LLMs Misfire Despite Knowing the Answer. arXiv:2605.22007, 2026.

Zellers, R. et al. HellaSwag: Can a Machine Really Finish Your Sentence? *ACL*, 2019. arXiv:1905.07830.

Zou, A. et al. Representation Engineering. arXiv:2310.01405, 2023.

A Full proof of Theorem 1

Write $M^x = M^0 + E(x)$ and $m_k^x = \frac{1}{d}\operatorname{Tr}((M^x)^k)$.

Step 1 (ensemble convergence). By Assumption 1(i) and free-probability self-averaging for rotationally-invariant ensembles with bounded fourth moments (Marčenko & Pastur, 1967; Bai & Silverstein, 2010), the empirical spectral distribution of M^0 converges almost surely to μ_l, so $m_1^0 \to \int t\, d\mu_l$ and $m_2^0 \to \int t^2\, d\mu_l$ are deterministic constants, and the spectral edge is almost surely bounded, giving $\|M^0\|_{\mathrm{op}} = O(1)$.

Step 2 (direct moment perturbation). Using $|\operatorname{Tr} A| \le \operatorname{rank}(A)\,\|A\|$ for the first moment and $\operatorname{rank}(2M^0E + E^2) \le 2\operatorname{rank}(E)$ for the second,

$$
\begin{aligned}
|m_1^x - m_1^0| &= \frac{1}{d}|\operatorname{Tr} E| \;\le\; \frac{1}{d}\operatorname{rank}(E)\,\|E\|, \\
|m_2^x - m_2^0| &= \frac{1}{d}|\operatorname{Tr}(2M^0E + E^2)| \\
&\le \frac{1}{d}\,2\operatorname{rank}(E)\left(2\|M^0\|\,\|E\| + \|E\|^2\right).
\end{aligned}
$$

By Assumption 1(ii), $\operatorname{rank}(E) = o(d)$ and $\|E\| = O(1)$, and $\|M^0\| = O(1)$ from Step 1, so both right-hand sides are $o(1)$. Hence $m_1^x \to m_1^0$ and $m_2^x \to m_2^0$.

Step 3 (assembly). $\mathrm{PR}(J_l(x))/d = m_1^2/m_2 \to (\int t\,d\mu_l)^2 / \int t^2\,d\mu_l =: c_l$, independent of x. □

Necessity of both conditions. Take $M^0 = I$, so $\mathrm{PR} = d$. A rank-one perturbation producing a single outlier eigenvalue at $\sqrt{d}$ has $\operatorname{rank} E = 1 = o(d)$ and shifts the empirical spectral distribution by $1/d \to 0$ in Kolmogorov distance, yet m_2 gains $(\sqrt{d})^2/d = 1$, so $\mathrm{PR} \approx d \cdot 1/2 = d/2$: the rank-only hypothesis fails. A full-rank perturbation of bounded norm shifts the Marčenko–Pastur bulk and changes m_1, m_2, so the norm-only hypothesis fails. The reason the proof bounds the moments directly rather than through distributional closeness is that Kolmogorov closeness does not control the first and second moments without uniform integrability. An outlier carrying $O(1)$ mass at a large eigenvalue moves m_2 while leaving the histogram essentially unchanged, and the operator-norm bound is exactly what forecloses such an outlier.

B THE CONTRACTION BRIDGE TO SAMPLING DETECTORS

The participation ratio characterizes the bulk spectrum, whereas sampling detectors read the dispersion of a finite cloud of generations. The bridge carries one further assumption, stated plainly. Provided the K-sample perturbations do not systematically align with class-specific outlier eigenvalues, input-invariant bulk geometry implies invariant local dispersion of the K-sample cloud around the endpoint. Viewing the residual stream as a forward-Euler discretization of a contractive dynamical system (Lu et al., 2018), the samples then contract to a sharp, stable output distribution regardless of correctness, so a sampling-based output detector such as semantic entropy is blinded for the same reason as a geometric one, under the same noise condition as Corollary 1. The empirical dispersion null of Appendix F.1 is consistent with the finite-sample bridge (non-separation at $n = 50$ per class), and contractivity is established empirically (spectral radius 0.829).

C EXPERIMENTAL DETAILS

Models. Qwen2.5 {0.5B, 1.5B, 3B, 7B, 14B, 32B}; Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B; Gemma-2 {2B, 27B} (the SAE analysis additionally uses Gemma-2-9B); Mistral-Small-24B; BF16, deterministic kernels, random seed 2026, unquantized. *Data.* A TruthfulQA-derived set used as an indivisible unit, and a six-dataset pool of 7,710 items (3,000 for the refusal analysis) spanning MMLU, ARC, HellaSwag, TruthfulQA, OpenBookQA, and CommonsenseQA in forced-choice / short-answer format; TriviaQA serves as the genuine-uncertainty positive control. *Probe.* Logistic probe on the commitment-band last-token residual, entity-grouped folds; features are the last-token states at fixed relative depths 0.68, 0.79 and 0.86, chosen once on Qwen-1.5B (817 TruthfulQA items, about 10.6% of the pool) and identical for every model. *Calibration and labels.* Confidence is temperature-scaled per model; a confident error is a wrong answer with calibrated confidence above τ; trap and ordinary (gap) labels follow the misconception-lure versus knowledge-deficit distinction. *Correctors.* Rank-8 LoRA on the attention projections per regime, trained category-held-out and merged into the weights; in the router evaluation of Section 9 the two per-regime adapters (query and value projections) are trained in-run on the train split and switched per item rather than merged; the commitment-layer steer adds a gold-direction vector at the selected layer with a per-size α schedule. *Router.* A regime classifier on the commitment-band features with leave-one-source-out evaluation and a source-identity-accuracy bound. *Power.* POWERED claims state n and a confidence interval; $n \approx 3$–48 is exploratory by default; de-confounding is a pre-run gate.

D ADDITIONAL RELATED WORK

Confident error, entropy, and supervised probes. Simhi et al. (2025a) identify CHOKE, high-certainty hallucinations on questions the model otherwise answers correctly, and show that certainty-based mitigation degrades on these cases while a probing-based mitigation outperforms it. Phillips et al. (2026) independently identify a model-dependent low-entropy failure regime and show that combining entropy with a supervised correctness probe improves selective prediction. Ma et al. (2025) propose an energy-based score on penultimate-layer logits. We complement these results with a structural account, under an idealized model, of why a class of unsupervised macroscopic endpoint-statistic readers is driven toward chance (Theorem 1, Corollary 1); a direct test of the supervised probe on precisely the errors semantic entropy misses; and regime-matched correction rather than abstention alone, which is necessary because the wrong intervention actively harms (Section 7).

Steering and dynamics. Inference-time intervention (Li et al., 2023), representation engineering (Zou et al., 2023), and contrastive activation addition (Rimsky et al., 2024) steer activations toward truthfulness. We use a commitment-layer steer as one actuator and report, as a negative result, that it is non-monotone in scale. Finally, viewing transformer depth as a forward-Euler discretization of a continuous dynamical system (Lu et al., 2018) supplies the contractive-dynamics framing for our dispersion null and sampling bridge, and work on Jacobian eigentracking (Fernando & Guitchounts, 2026) studies the same operator we analyze, though for spectral geometry rather than input-invariance-as-blindness.

Mechanistic interpretability. A complementary line decomposes activations into sparse, often human-interpretable features via dictionary learning (Bricken et al., 2023; Templeton et al., 2024) and traces multi-step computation through attribution graphs (Lindsey et al., 2025). That program asks *what* features and circuits a model uses, whereas our question is whether the *macroscopic spectral geometry* of the committed representation can, without supervision, separate confident error from correct. A negative answer for spectral readouts neither competes with nor is contradicted by the existence of interpretable features. A feature for 'unsure' or 'unknown-entity' may exist and be readable with supervision even while the coarse spectral geometry remains class-invariant. We show (Appendix F.4) that an unsupervised SAE readout inherits this blindness.

E EXTENDED IDEALIZED INVARIANCE AND EMPIRICAL SCOPE

Assumption 1(i) formalizes the idea that the bulk of a wide layer's transition is set by its weights, whose spectrum self-averages;[1] (ii) formalizes two distinct facts about attention and gating: that they route through comparatively few directions (low rank) and that they do so with bounded gain (bounded norm).[2] We emphasize, and the proof requires, that *both* parts of (ii) are substantive. Neither implies the other, and dropping either breaks the conclusion.

Both conditions are necessary. The two parts of Assumption 1(ii) are not redundant. Take $M^0 = I$, so $\mathrm{PR} = d$. A rank-one perturbation that places a single outlier eigenvalue at $\sqrt{d}$ has rank $1 = o(d)$, and the empirical spectral distribution moves by only $1/d$ in Kolmogorov distance, yet the second moment gains $(\sqrt{d})^2/d = 1$, so $\mathrm{PR} \approx d \cdot 1/2 = d/2$: the participation ratio halves. Rank control alone therefore fails, because a low-rank perturbation of large norm can move the spectral moments by $O(1)$ while leaving the histogram essentially unchanged. Symmetrically, a full-rank perturbation of bounded norm shifts the Marčenko–Pastur bulk[3] and changes m_1, m_2, so norm control alone

[1] Self-averaging is the random-matrix phenomenon that, as a matrix grows wide, the histogram of its eigenvalues stops fluctuating from sample to sample and converges to a fixed deterministic shape (here μ_l). A rotationally-invariant ensemble with bounded fourth moments is the standard idealization of a generic wide weight matrix for which this holds. Such an ensemble has statistics that look the same in every basis, with entries that are not heavy-tailed.

[2] Two independent controls. Low rank says attention and gating touch comparatively few directions relative to the model width; bounded norm says they do so with limited gain, amplifying no direction without bound. The result needs both, as the next paragraph shows. Low rank alone permits a single large spike that halves the participation ratio, and bounded norm alone permits a broad shift of the whole spectrum.

[3] The Marčenko–Pastur law is the random-matrix analogue of the central limit theorem. It gives the limiting eigenvalue distribution of a large random covariance-type matrix. A full-rank perturbation shifts this bulk, which is why a bounded-norm change can still move the spectral moments even though it introduces no large spike.

fails as well. The proof bounds the moments *directly* rather than through spectral-distribution closeness precisely because distributional closeness does not control the moments without a uniform-integrability condition that the operator-norm bound supplies. This is the mathematically delicate point of the result and is treated in full in Appendix A.

F EXTENDED SEMANTIC ENTROPY AND CONFIDENT ERROR

The endpoint-blindness conclusion does not rest on the single network where we measure the participation ratio. Independent unsupervised readers corroborate it across all three families. A dispersion readout on Qwen-0.5B/1.5B/3B and Llama-3B, and a Gemma-Scope reconstruction-error readout on Gemma-2-2B and Gemma-2-9B, all fall within an equivalence-to-chance band at the commitment band, far below the supervised probe on the same states, which separates the regimes.

This bound reaches beyond the geometric readers of the endpoint to the sampling-based detectors as well. Semantic entropy reads the dispersion of a cloud of samples rather than the endpoint geometry directly, but under contractive dynamics a confident error's samples contract to the same tight cloud as a correct answer's, so, under the conditions of Appendix B, they cannot disperse to flag the error. We give the contractive-dynamics argument in Appendix B and test it directly below.

The theorem thus predicts that unsupervised endpoint and sampling detectors are alike blind to confident error. We test the prediction four independent ways, three behavioral and one on internal features, each on the deployment-relevant contrast of confident-wrong against correct.

F.1 HIDDEN-STATE DISPERSION DOES NOT SEPARATE

If, as the contraction bridge argues, the sampled continuations of a confident error contract to the same tight cloud as those of a correct answer, then dispersion should not separate the two classes. It does not. For each item we draw $K = 16$ samples, take the commitment-band last-token state of each, and measure the cloud two ways, by total variance and mean pairwise cosine distance, matched on confidence. Both measures fall within the equivalence-to-chance band on all four models spanning two families. Qwen-0.5B (0.407, 0.409), Qwen-1.5B (0.479, 0.482), Qwen-3B (0.477, 0.489), and Llama-3.2-3B (0.487, 0.501) all sit near chance, the non-separation holding across both measures and convergent with the other three unsupervised readouts. We are careful about the strength of the claim. At $n = 50$ per class this is a failure to separate, not a proven null, so we do not assert the absence of a signal. What we report is a non-separation that holds across every model and measure we tested and matches what the contractive picture predicts.

F.2 SEMANTIC ENTROPY IS BLIND, AS A DOUBLE DISSOCIATION

We compute Farquhar semantic entropy (Farquhar et al., 2024) with bidirectional-entailment NLI clustering[4] (`roberta-large-mnli`, $k = 8$ samples). On TruthfulQA, an adversarial misconception set in which the errors are confident, semantic entropy does not separate wrong from correct. Across twelve models and four families the AUROC ranges from 0.476 to 0.563, and every 95% confidence interval has an upper bound below 0.65 (Table 1, Figure 3). Scale does not rescue it. Isolating confident error from the genuine ignorance that semantic entropy legitimately detects, the confident-error-specific AUROC scatters around 0.5 ($[0.487, 0.577]$, with values below 0.5). This is consistent with the idealized theorem's exact 0.5 once two effects are admitted: finite-sample noise, and a $\leq 10\%$ label/refusal contamination (one-signed, inflating the AUROC by at most about 0.05). The mild rise of the *aggregate* with scale is the ignorance arm of the dissociation that semantic entropy correctly catches, not recovered sensitivity to confident error. Operationally, semantic entropy is a near-chance, useless detector of confident error; its AUROC point estimates are bounded below ≈ 0.60, far beneath the supervised probe (0.73–0.85, Section 6) and any deployable threshold. We

[4] Semantic entropy quantifies uncertainty over *meanings* rather than tokens. The model is sampled k times; a natural-language-inference model groups the generations into clusters of mutually-entailing meaning; the entropy is taken over the cluster probabilities. Dispersed meanings give high entropy (genuine uncertainty); agreeing samples give low entropy. It is among the strongest black-box hallucination detectors, which is exactly why its blindness to confident error is informative rather than a weak-baseline artifact. That blindness occurs precisely where the samples agree on a single wrong meaning.

frame this blindness as an operational equivalence to chance rather than a literal 0.5. Formally, it is an equivalence claim, established by two one-sided tests. The claim is that the confident-error AUROC lies within a deployment margin of chance, $|\mathrm{AUROC} - 0.5| < 0.10$. It must also remain far below the supervised probe. This is the power-appropriate claim. At large n, a trivially small deviation from 0.5 can reach statistical significance without approaching usefulness. The robust, deployment-relevant statement is therefore that the detector is bounded into a useless band, not that it is exactly 0.5. Two facts make 0.5 the correct reference and the observed wobble above it benign. First, the idealized theorem predicts exactly 0.5, and by the central limit theorem the finite-sample AUROC concentrates on its population value as n grows, so a small finite-n excursion above 0.5 is consistent with sampling noise rather than signal. Second, the known contamination is *one-signed*. Mislabeled refusals and genuine-ignorance items carry high semantic entropy that the detector legitimately catches, so they can only *inflate* the confident-error AUROC. A contamination fraction ε shifts it by at most $O(\varepsilon)$, and the $\leq 10\%$ refusal and label-noise rate we observe is sufficient on its own to account for the 0.50–0.55 excursion. Therefore, the measured values are, if anything, an *upper bound* on the clean confident-error AUROC, which lies even closer to chance. No additional experiment is needed to make the equivalence claim conservative. On TriviaQA, where the errors are genuine recall failures, the same pipeline scores $\mathrm{AUROC} = 0.774$. The contrast is a double dissociation:[5] a single method, on a single contrast, works when the uncertainty is genuine and fails when the error is confident. Because the dataset determines the regime (TruthfulQA contains no genuine ignorance), the contrast it supports is precisely confident-wrong against correct, with the ignorance arm supplied by the TriviaQA control rather than by a within-set split. The dissociation is robust to the stylistic-geometry confound by which false answers may lie closer to the query (Marín, 2026), because semantic entropy clusters by entailment rather than by cosine distance.

The confident-wrong subset was identified in part through the model's own output behavior, which raises a circularity concern. A detector tested on items its own pipeline helped select could appear to separate them for trivial reasons. To address this, we selected the forced-choice A/B margin as an independent confidence axis, a different task from open-generation semantic entropy that labels confident-wrong items without reference to the sampling signal. On this independent split, semantic entropy still fails to separate confident-wrong from correct. The AUROC is 0.481, 0.468, and 0.469 on Qwen-0.5B, 1.5B, and 3B ($n = 125$, 134, 131 confident-wrong), pre-registered before the 1.5B and 3B runs.

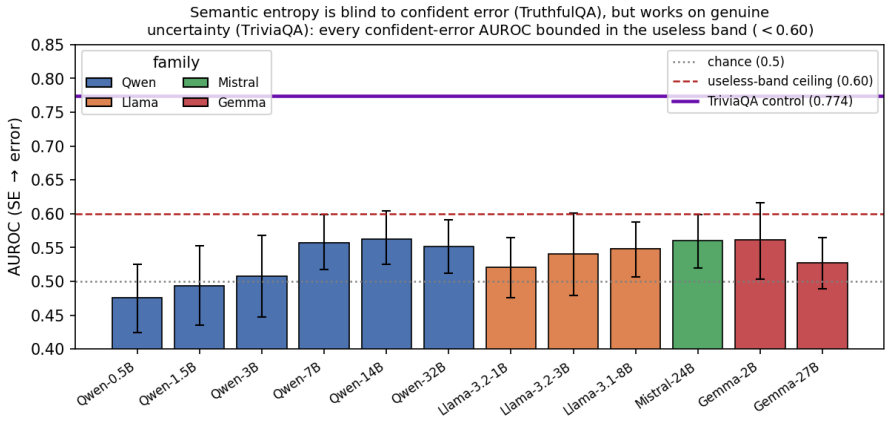

Figure 3: The double dissociation. Semantic entropy does not separate confident-wrong from correct on TruthfulQA (bars; all 95% CI upper bounds below 0.65, point estimates below 0.60), while the same pipeline scores 0.774 on TriviaQA (control line), where uncertainty is genuine. Bars show all twelve models in Table 1.

[5] A double dissociation is the strong evidential pattern in which one method, applied to one kind of contrast, succeeds in one condition and fails in another, with only the condition changed. Here semantic entropy works on genuine-uncertainty errors (TriviaQA) and fails on confident errors (TruthfulQA); because the pipeline is identical, the failure is a property of confident error, not a broken measurement.

F.3 SEMANTIC ENTROPY IS INVERTED ON AN INDEPENDENT LABEL

Joining open-generation semantic entropy to a forced-choice error label produces a suggestive inversion. The AUROC is 0.443, with a 95% CI of $[0.375, 0.516]$, and confident-wrong answers carry *lower* median semantic entropy (1.49) than correct ones (1.67). A sampling detector rates the confident error *more* trustworthy than the correct answer. We read this cautiously. The CI crosses 0.5, so the inversion is suggestive rather than separated from chance, and TruthfulQA's lures are engineered to be high-probability under pretraining statistics (Kalai et al., 2025), so the effect may partly reflect adversarial dataset construction and should not be over-generalized.

F.4 AN UNSUPERVISED SPARSE-AUTOENCODER READOUT IS ALSO BLIND

The blindness extends from the macroscopic spectral geometry to dictionary-learning readouts, *across two model scales*. At the commitment band of Gemma-2-9B, an *unsupervised* Gemma-Scope (Lieberum et al., 2024) sparse-autoencoder's reconstruction error, magnitude-controlled, does not separate confident error from correct (AUROC from 0.475 to 0.590 across three layers, near chance ($|\mathrm{AUROC} - 0.5| < 0.10$) even at layer 24, where it reaches significance (0.590, permutation $p = 0.0005$) without reaching usefulness), whereas a supervised probe on the same states does (0.65–0.70). We replicate this on Gemma-2-2B with the size-matched Gemma-Scope-2B SAE at powered $n = 559$ (260 confident-wrong, 299 correct): the unsupervised reconstruction error is again at chance ($\mathrm{AUROC} = 0.497$, 95% CI $[0.448, 0.545]$, permutation $p = 0.56$; mean reconstruction error 81.19 vs. 81.18 for confident-wrong vs. correct), while the supervised commitment-band probe separates these same regimes (Section 6). The unsupervised dictionary-feature readout thus inherits the endpoint blindness at both scales. There is one caveat we cannot yet close. Both public SAEs are base-model-trained but applied to instruction-tuned models, and a base-on-base control was not run (the base Gemma-2-2B weights are access-gated).

Taken together, the four measurements confirm the prediction. The strongest available unsupervised tools do not see confident error, and the signal is not at the endpoint.

G EXTENDED EXPERIMENTAL SETUP

Evaluation models. Three evaluation components sit outside the models under study and are held fixed across the ladder. Semantic-entropy clustering uses a natural-language-inference entailment model, `roberta-large-mnli`, applied bidirectionally to group $k = 8$ sampled generations by meaning (Farquhar et al., 2024); we deliberately use an entailment judge rather than cosine distance so the measurement is immune to the stylistic-geometry confound (Marín, 2026). Calibrated confidence is produced by per-model temperature scaling (Guo et al., 2017). The detector itself is a logistic probe trained on commitment-band features (Section 6); it is the only learned evaluator, and it is trained and tested with entity-grouped folds so that no entity is shared across the split.

H EXTENDED SUPERVISED DETECTOR

The bound forecloses unsupervised endpoint detection but explicitly permits a supervised probe. We train a linear probe on the commitment-band last-token residual to flag confidently-wrong risk, using entity-grouped folds so that the same entity never appears in both train and test, with per-model error counts of 85 to $1{,}715$ across the models (the most accurate larger models contribute fewer). *Crucially, the supervised read generalizes across scale and across all four model families, and, unlike the blind unsupervised endpoint geometry of Section 3, it strengthens with scale across the Qwen ladder, precisely where an endpoint detector fails.* On the verified local ladder the probe reaches AUROC 0.731 (95% CI $[0.715, 0.746]$) on Qwen-0.5B, 0.779 $[0.760, 0.797]$ on Qwen-1.5B, 0.762 $[0.738, 0.787]$ on Qwen-3B, 0.802 $[0.760, 0.843]$ on Qwen-14B, and 0.821 $[0.771, 0.865]$ on Qwen-32B (Figure 5). Across families it clears the confound floor on every model and is strongest on Llama-8B at 0.849, with Gemma-27B at 0.775. Mistral-24B (0.727) is reported separately as a marginal cross-family probe, its high accuracy yielding few, hard errors. This is a direct double dissociation against the semantic-entropy blindness that *deepens* at those same scales. The Qwen-ladder detector AUROCs (0.5B/1.5B/3B) are POWERED: pre-registered (a frozen, one-shot decision rule), with real confident-wrong n (e.g. $n = 1715$ at 0.5B), tight CIs, cleared confound floors, and

an independent re-implementation. The achieved power is overwhelming. At $n = 1715$ the 0.5B AUROC sits $\approx 21\sigma$ above the dataset-confound floor ($> 99\%$ power to reject AUROC=floor), with the rest of the floor-clearing ladder 16–20σ above. The cross-family floor-clearers (Llama-8B, Gemma-27B, Qwen-14B/32B) are corroborated by the two validations below, which are likewise POWERED.

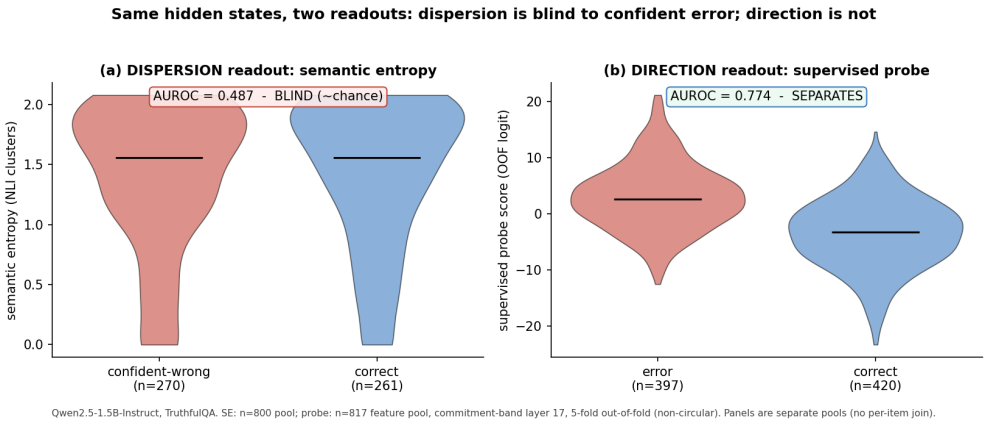

Figure 4: The same hidden states of Qwen2.5-1.5B-Instruct (TruthfulQA), read two ways. (a) A *dispersion* readout (semantic entropy over NLI-clustered samples) does not separate confident-wrong from correct (AUROC $= 0.487$; confident-wrong defined by an independent forced-choice margin, so the split is non-circular). (b) A *direction* readout, a supervised commitment-band probe scored out-of-fold, separates error from correct (AUROC $= 0.774$). Truth lives in a direction the dispersion readout cannot see; this is the bound (Section 3) and its escape hatch (this section) in one picture. Separate item pools of the same model. POWERED.

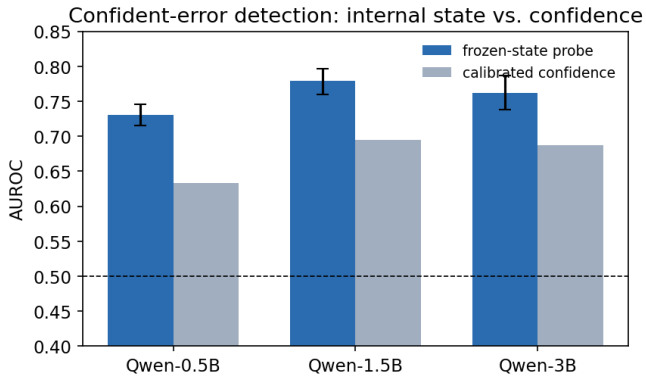

Figure 5: Supervised confident-error detection against calibrated confidence on the verified local ladder, with 95% intervals. The internal-state read clears the confidence baseline at every scale.

Table 4: De-confound battery. Each test isolates a plausible nuisance variable; the probe passes all four.

Test	Result	Reading
Length	full 0.783 vs. norm-only 0.591; $r = 0.02$	length is not the driver
Entity leakage	entity-held-out gap $+0.051$	small
Prevalence	balanced-prevalence AUROC 0.785	invariant
Structure	feature-shuffle AUROC $\to$ 0.502	signal is in feature structure

The SE-blind result of Section 6 is a direct confirmation of the supervised/unsupervised dissociation: the read recovers exactly the confident errors the strongest black-box detector structurally cannot see. With detection in hand, the next question is whether one correction serves both error regimes.

I Extended corruption and rescue

The finer mechanistic claim, that rescue specifically requires the dominant subspace as isolated by the in-subspace control, replicates cleanly only on Qwen-1.5B and is band- and α-sensitive on Gemma-2B and Llama-3.2-3B, so we scope that mechanism to Qwen and mark it exploratory.

Table 5: The subspace asymmetry. With the dominant variance modes removed, corruption survives (intervention effect vs. the isotropic control) across 3/3 families, while rescue collapses in every condition. The finer rescue-needs-the-dominant-subspace mechanism is Qwen-only (scoped exploratory).

Family	Corruption (orth. vs. isotropic) orthogonalized	isotropic	Rescue (all conditions)
Qwen-1.5B	0.92	0.35	≈ 0
Gemma-2B	0.73	0.37	≈ 0
Llama-3.2-3B	0.96	0.08	≈ 0

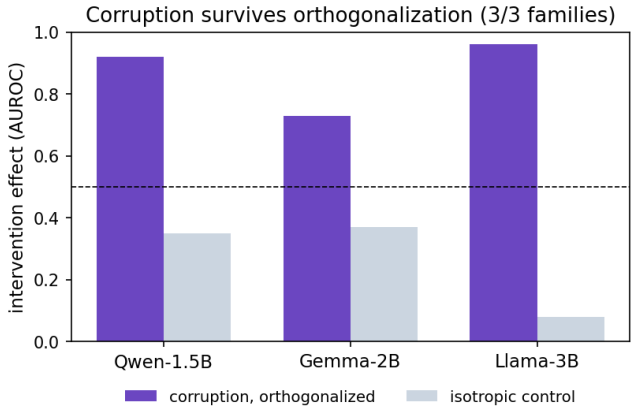

Figure 6: Corruption rides the specific direction and survives orthogonalization against the dominant variance modes (versus the isotropic control), three families out of three; rescue collapses in all conditions (not shown).

J Extended integrated Governor

With the mechanism in hand, two questions remain before the system is whole: whether the failure survives scale, and whether the parts assemble into a single runtime layer. We take them in turn. *The basin persists.* Before assembling the system we ask whether the confident-error basin is a small-model artifact that scale would dissolve. It is not. The blindness bound is architectural rather than scale-specific, since $\mathrm{PR}(J_l)$-invariance predicts the null at any width, and the basin itself persists across 1.5B to 32B: it remains reachable by activation patching (window-patch flips the answer at roughly 0.95 to 1.0 at every scale), but it is not reachable by steering, and the barrier to entry steepens with scale rather than the basin vanishing (Akarlar, 2026; Cherukuri & Varshney, 2026). The failure the Governor addresses does not disappear as models grow, which is exactly what justifies a deployable governance layer.

Table 6: Within-family deployment comparison: *all* models scored by letter-choice (the deployed operating mode), identical $n = 150$ items/benchmark, seed-matched. *+abs* = the small model answers only its confident half (top-1 minus top-2 letter-logit margin; 50% coverage); the large model answers all. Bold = the abstaining small model reaches or passes its within-family large model. The scale jump is 13–47×.

	Llama (3B→70B)			Qwen (1.5B→32B)			Gemma (2B→27B)		
Benchmark	sm	+abs	lg	sm	+abs	lg	sm	+abs	lg
MMLU (recall)	.57	**.81**	.80	.62	**.81**	.79	.65	**.83**	.74
ARC (reason)	.71	.87	.91	.74	.92	.95	.71	**.92**	.91
OpenBookQA	.74	**.91**	.91	.71	.87	.93	.65	.87	.92
CommonsenseQA	.73	**.87**	.85	.75	**.93**	.87	.73	**.87**	.79
HellaSwag	.39	.41	.83	.35	.37	.83	.51	.56	.75
WinoGrande	.59	.65	.75	.55	.59	.81	.61	.64	.72
alphaNLI	.75	.88	.93	.65	.83	.94	.79	**.95**	.89

Deployment: a small model that knows when it knows. Selective answering is most valuable when a small model is asked to stand in for a much larger one. Table 6 compares each small model against the *next model up in its own family*, a 13–47× scale jump, all scored by letter-choice, the realistic deployed mode. A clean pattern emerges: *at 50% coverage the abstaining small model matches or passes its within-family large model (which answers every item) on the knowledge and reasoning benchmarks* (MMLU and CommonsenseQA in all three families, ARC/OpenBookQA/alphaNLI in one each), so the half it is confident about is answered at large-model accuracy despite a 13–47× scale gap. The exception is the *completion* benchmarks (HellaSwag, WinoGrande), where the large model stays ahead. The reason is the two-regime structure that runs through this paper. Where the small model *has* the knowledge, its confident answers are usually right, so keeping the confident half lifts it to large-model accuracy; where it *lacks* the knowledge (facts it never stored, multi-step reasoning it cannot do), confidence-filtering cannot manufacture the answer, and the correct move is to *defer*.

This reframes what a deployable small model is for. In a safety-critical, air-gapped, or regulated setting the operative quantity is *calibrated coverage* rather than raw accuracy. A model that answers what it knows at high accuracy and says *"I don't know"* on the rest is more deployable than a larger model that is confidently wrong, because the failure it removes (a fluent, unflagged error) is the one that does damage. A confidence signal supplies exactly that; the abstention column uses its simplest form, the top-1 minus top-2 letter-logit margin, trading coverage (50% here) for trustworthiness at a tunable operating point. The learned detector and the full Governor, reported in Sections 6 and 9, go further: matched-coverage refusal and regime-routed correction.

Two lenses, and the eval is part of the result. We scored every model two ways. *Letter-choice* (above) measures *deployed* accuracy, the answer produced through the chat interface. *Loglikelihood* measures *latent* knowledge, whether the model's probability mass favors the right option, and it is the native space of our confidence signal. The two diverge informatively. For large instruction-tuned models letter-choice far exceeds loglikelihood (e.g. Qwen-32B MMLU $0.54 \rightarrow 0.79$), because instruction-tuning is what *accesses* latent knowledge, so the gap itself measures access versus knowledge. They even *invert* by task: HellaSwag is sentence-*completion*, native to loglikelihood, and forcing it into a letter prompt drops a small model from 0.71 to 0.35. This is *suggestive of* more than abstention alone. Coupled with the Governor's corrective adapter and *selective escalation* (route only the detector-flagged residual to a frontier model, not every query), a governed small model could serve the majority of traffic at high reliability and pay frontier-scale cost only on the minority it genuinely cannot handle. The within-family parity that abstention already reaches on knowledge tasks is the first, conservative piece of that efficiency argument.

K Extended negative results

We report the system's limits as results, because each one fixes a component of the operating envelope and tells a deployer what to rely on. *Steering is non-monotone in scale.* The commitment-layer steer corrects $+0.233$ in the right regime at $\leq$3B and is effective only at the commitment layer, but its net effect across regimes is $+0.158$ $[0.067, 0.258]$ at 1.5B, -0.045 $[-0.075, -0.02]$ at 3B, and $+0.179$

$[0.065, 0.311]$ at 14B. This is a robust mid-scale failure with recovery at the large end, which refutes a simple attenuation story; at scale the pipeline therefore relies on adapters and refusal. *Separability is confounded at small scale.* As Section 7 establishes, perfect trap-versus-gap separability is, at small scale, an artifact of dataset-identity leakage rather than a regime signal (regime-driven is False under the source-crossed control); crucially, this leakage scales *inversely* with model size and gives way to the genuine, capacity-emergent regime representation, which strengthens at the scales that matter. We therefore report raw separability as a confounded artifact, not a contribution, and rest the routing claim on the leakage-controlled, capacity-emergent signal. *There is no capacity cliff.* A requisite-variety stress test that raises demand finds the per-item flux observable rising from 4.42 to 17.15 against only a gentle accuracy decline, which removes a capacity-bottleneck reading as a causal mechanism in this regime (exploratory). *The substrate dynamics are generic, and dramatic dynamical signatures do not survive.* We also asked whether confident error carries a distinctive *dynamical* signature in the residual-stream flow, as basin and attractor accounts might suggest (Cherukuri & Varshney, 2026; Akarlar, 2026). Under a whitened, fluctuation-relative metric the dynamics are contractive (spectral radius 0.829), consistent with viewing transformer depth as a contractive discretized flow (Lu et al., 2018); but the macroscopic observables do not discriminate confident error: the rotational fraction (about 0.76) is universal across vendors and at chance (AUROC 0.44), and a fluctuation–response excess is not a genuine probability current. Two earlier, more dramatic readings (a non-equilibrium signature and a gyroscopic-stabilization effect) did not survive internal audit and are retired as raw-metric artifacts. We therefore *bound* dynamical-signature accounts at the macroscopic level. The basin phenomenology is real, but the discriminating signal is representational rather than dynamical, and the Governor depends only on the former. These substrate observables are exploratory ($N = 1$).

L Discussion

The organizing lesson of this paper is that confident error is *governable*, not detectable-and-filterable. That distinction is worth stating sharply because it redirects the research program.

Detection is structurally, not just empirically, hard. The field has treated the difficulty of catching confident hallucination as an engineering gap, one more unsupervised score away from being closed. Our theorem says the gap is, for unsupervised readers of the macroscopic endpoint statistic, closed in principle. The geometry these detectors read is input-invariant, so reading it more cleverly cannot separate a confident error from a correct answer. A deep well looks like a deep well whether the answer is fact or fabrication. The productive response is a different object of measurement, not a better endpoint score. It is a supervised read of the *path* the model takes to its answer, together with the recognition that detection alone is insufficient once errors fall into regimes that demand different, mutually harmful corrections. The Governor is the architecture this analysis forces, and the field's effort is better spent there than on the bounded problem.

Why confident error erodes trust, and why refusal restores it. The deeper reason to catch a confident error is that it is the one failure mode that defeats trust calibration rather than merely producing a wrong answer. A low-confidence error is self-flagging; its hesitation invites scrutiny and teaches the operator the correct lesson, that the system's own uncertainty is the cue to verify. A confident error teaches the opposite. It presents the same fluent, assertive surface as a correct answer, and Section 3 shows that the readable endpoint geometry is identical as well, so the operator is left no signal, at the surface or in the geometry, on which to withdraw reliance. Repeated exposure therefore does something worse than cause isolated mistakes; it trains the operator that fluency and assertiveness mark reliability, which is exactly the cue the next confident error exploits. The failure thus corrodes the operator's trust model itself, and calibrated trust is among the load-bearing requirements that the human-machine teaming systems-engineering literature names for a usable autonomous system (McDermott et al., 2018).

A governed refusal is, on this reading, not a degradation of service but a transparency act. By reading residual risk from the model's own internal state and declining when that risk is high, the system communicates the boundary of its competence at the moment of use, the function that situation-awareness-based agent transparency identifies as central to human-autonomy teaming, namely conveying an agent's confidence and the limits of its knowledge so the operator relies on it appropriately (Chen et al., 2018). This account treats the agent's uncertainty as an explicit

transparency element (Stowers et al., 2016), and earlier work has evaluated how such transparency affects human performance in these teams (Stowers et al., 2020). The Governor supplies that signal through an unconventional channel. Transparency is usually drawn from the agent's self-report, yet this paper shows self-report is precisely what fails on confident error, since the model's own calibrated confidence is among the baselines the supervised probe beats. Reading the regime from the representational substrate therefore furnishes the transparency signal from a source the answer's surface conceals, and one that does not depend on the model's own compromised confidence. We claim this as enabling calibrated reliance, not as a demonstrated teaming gain; the present work measures detection, routing, and refusal rather than operator situation awareness or team performance, and the teaming implications are developed in companion work.

What the theorem licenses, and what it does not. The result is easy to over-read, so we are explicit. It does *not* say hallucination is undetectable; it says one specific, popular class of detectors is bound to chance, and the load-bearing reason is the supervised/unsupervised boundary. A supervised probe reads sub-macroscopic, off-spectrum structure the theorem does not constrain, which is exactly why the internal-state detection literature works and why our own probe works. There is no contradiction. The cleanest way to place the contribution alongside the recent statistical-inevitability results (Kalai et al., 2025) is that hallucination may be *inevitable*, but it is *governable*. Inevitability is a statement about the training objective; governability is a statement about runtime, and the two are complementary rather than opposed.

The corruption/rescue asymmetry is a safety result, not only a mechanism. That corruption is carried by a specific direction while rescue requires the dominant subspace cuts two ways. Defensively, it explains why steering toward truth is fragile and argues for weight-merged adapters over runtime steering in deployment. Adversarially, it warns that pushing a model *into* confident error is cheap (a single direction suffices), so the same geometry that makes honest correction hard makes dishonest corruption easy. A governance layer that reads regime and refuses is, on this reading, a defense against an asymmetry that favors the attacker.

What scales, and what does not. The picture across scale is not uniform, and the honest version is more useful than a flat "it works." The steering actuator fades and reverses with size, so it is not the lever to lean on. The regime representation, by contrast, is *capacity-emergent*, sharpening as models grow, so the routed, refusal-first governor improves on exactly the models that matter for deployment. The endpoint-blindness bound, being architectural, holds at every scale we can reach. A deployer should therefore build on the detector, the router, and refusal, and treat steering as a small-model convenience. The phenomenon connects naturally to the broader view of confident commitment as an absorbing basin in the model's dynamics; we touch that framing only lightly here, as the present claims rest on the random-matrix argument and its empirical validation rather than on any larger theory.

M FUTURE DIRECTIONS

Three directions follow directly. The first is open-generation deployment: bridging from forced-choice to long-form generation, where the unit of error is the entity within free text, connects our governor to the long-form entity-detection literature and is the natural path to production use. The second is the boundary of the theorem itself. Real attention can in principle violate the bounded-norm condition of Assumption 1, and characterizing the inputs, if any, on which it does would both sharpen the bound and reveal where an unsupervised detector could, after all, succeed. The complement of blindness is a detector waiting to be built. The third is to characterize geometrically *what* the supervised probe reads. The theorem tells us the signal is off the macroscopic spectrum, but not yet which sub-macroscopic directions carry it, and identifying them would turn an empirical detector into a principled one.

These are pursued in companion work; they are forward pointers, not claims on which the present results depend. The geometry underlying the supervised signal (what the probe reads, and why the confident error is representationally separable) is treated in a companion paper; the confident-wrong state is further characterized as a causal, on-manifold attractor (its persistence to 72B and its architecture-dependent barrier) in related work in preparation; and the implications for how a

human operator relies on, calibrates trust in, and is supported by a governed system are developed in a companion paper on governed human–machine teaming.

N GLOSSARY OF TECHNICAL TERMS

For readers outside random-matrix theory or uncertainty estimation, we gloss the technical terms used in the main text.

Jacobian. The Jacobian J_l is the matrix of first derivatives of a layer's output with respect to its input; to first order it captures how the layer stretches and rotates the residual stream around the current activation. Its singular values σ_i are the per-direction stretch factors, and $M = J_l^\top J_l$ collects their squares, so the geometry any endpoint detector reads is ultimately some function of this spectrum.

Participation ratio. The participation ratio counts how many directions a transformation effectively uses. If it stretches space comparably in k directions and negligibly in the rest, the ratio is about k; if it concentrates all of its action in a single direction, the ratio is near 1. The form $(\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$ makes this precise: it is large when the squared singular values are spread evenly and small when one dominates, which is why it is the cleanest single-number summary of the local geometry an unsupervised detector would compute.

AUROC. AUROC (area under the receiver-operating-characteristic curve) is the probability that a randomly drawn positive case (here a confident error) receives a higher detector score than a randomly drawn negative one (a correct answer). 1.0 is perfect ranking, 0.5 is a coin flip with no discriminating power, and below 0.5 means the score is systematically inverted. It is threshold-free and insensitive to how rare errors are, which is why we report it throughout.

Semantic entropy. Semantic entropy quantifies uncertainty over *meanings* rather than tokens. The model is sampled k times; a natural-language-inference model groups the generations into clusters of mutually-entailing meaning; the entropy is taken over the cluster probabilities. Dispersed meanings give high entropy (genuine uncertainty); agreeing samples give low entropy. It is among the strongest black-box hallucination detectors, which is exactly why its blindness to confident error is informative rather than a weak-baseline artifact. That blindness occurs precisely where the samples agree on a single wrong meaning.

Commitment band. Empirically a transformer fixes its answer at an intermediate band, about 0.68–0.86 of relative depth, rather than at the output layer; after this band the prediction changes little. Reading the residual stream in this commitment band lets a supervised probe observe the *path* to the answer rather than the collapsed endpoint to which the unsupervised detectors of Section 3 are bound.

Selective prediction and risk–coverage. Selective prediction lets a system abstain rather than answer. A risk–coverage curve plots the error rate (risk) as the system answers a growing fraction of inputs (coverage); a good gate keeps risk low while covering more, and the area under the curve summarizes how well it separates safe from unsafe cases.

Anti-transfer. Anti-transfer is negative transfer between regimes, where a corrector tuned for one regime actively harms the other rather than merely failing on it. It is what upgrades "two regimes exist" into "the system must route": a single shared fix would damage half the cases.

McNemar's test. McNemar's test compares two classifiers on the *same* items by examining only the cases where they disagree and asking whether the wins are too one-sided to be chance; $p < 0.001$ means the router's edge over the best single fix is not a sampling fluke.

Capacity-emergent. A property is capacity-emergent if it appears only once a model is large enough. The genuine regime signal is absent at 0.5B, where the classifier instead exploits dataset identity, and strengthens with scale, so the routing signal we depend on is one that *improves* on the larger models that matter most for deployment.

Orthogonalization. Orthogonalizing away the top variance modes projects the activations onto the subspace left after the few highest-variance directions are removed. An effect that survives cannot be carried by those dominant directions alone; it must ride a more specific direction. We use it as a scalpel to ask which directions carry corruption versus rescue.

Low-rank adapter (LoRA). A low-rank adapter (LoRA) is a small set of trainable weight updates (here rank 8) added to an otherwise frozen model. It installs a regime-specific corrector without retraining or duplicating the full model, and merging it into the weights makes the correction permanent and free at inference.