Stable wrong beliefs evade every self-check: a literature synthesis on model self-knowledge and confident errors

W. Schulz

Nov 11, 2026

Abstract

A literature synthesis on whether language models know what they don't know. Self-derived signals (sampling disagreement, hidden-state probes, trained verbalized confidence) detect instability well but fall to about 0.55–0.67 AUROC on errors a model repeats. Catching those needs outside information. Includes six queued experiments; many sources are unreviewed 2025–26 preprints.

W. Schulz, NeuroGenesys, November 2026. A literature synthesis, not an experiment: the numbers are the cited authors' own and were not reproduced here, except where a NeuroGenesys study is named. Many sources are 2025–26 preprints, several unreviewed, and some were read only at abstract or search-snippet level or through secondary summaries; those are flagged where they appear. One frequently quoted figure that could not be verified (a gpt-5-thinking-mini abstention rate) is omitted, and arXiv 2602.13093 is flagged below as internally inconsistent. The comparisons with NeuroGenesys results refer to the selector-abstention and belief-repair papers. Research notes: `research_notes/Model self knowledge and confident errors/` *in the NeuroGenesys repository.*

Language models can tell what they don't know only in one sense: they detect their own *instability*. They cannot detect a wrong belief they hold steadily. The best training-free detector, semantic entropy, flags confabulations with an average AUROC of 0.79. Hidden-state probes reach 0.83–0.95 at the exact answer token. Both fall to roughly 0.55–0.67 on errors the model makes the same way every time. That is the band E1 measured on negated beliefs (0.54–0.55), and the band where verbalized confidence lands under social pressure (0.57). Inside the model, the "do I know this?" gate keys on *familiarity with the entity* rather than on *whether the answer is available*. A recognized name therefore switches off a default "can't answer" circuit even when the specific fact is missing. Training then makes it worse: binary-graded benchmarks and RL on verifiable rewards pay for every guess, so post-trained and reasoning models abstain far less and state high confidence even when wrong. Two training patterns reliably undo this. One attaches an abstain target to the model's *own* measured unknowns (R-Tuning, "I don't know" relabelling, unanswerable-data mixing). The other rewards a verbalized confidence with a proper scoring rule and thresholds it at deployment (RLCR, behaviorally calibrated RL). The second cuts in-domain ECE from 0.37 to 0.03 with no loss of accuracy. System guards add guarantees, but conformal methods certify *rates over a population*, not individual answers. Retrieval triggers, cascades and self-verification all inherit the model's blindness to its own confident errors. The guards that actually catch a stable wrong belief bring in *new information*: search-backed verification, an independent model family, structural priors, or density on a shared representation for out-of-distribution queries. For NeuroGenesys, the literature confirms the selector-abstention campaign's core result: change the scorer, and give abstention out-of-competence evidence. It tempers the strong form "never use the voter's own confidence". Disagreement across a model's own samples, its hidden states, and its RL-trained verbal-

ized confidence are all self-derived signals that work. They work on uncertain errors, not on stable ones.

Self-uncertainty signals measure instability, so they miss stable errors

Every training-free uncertainty method asks the model some version of "would you say this again?" Methods differ in how they ask and in what they cost. The Nature 2024 semantic-entropy study clusters about 10 sampled answers by bidirectional entailment and computes entropy over meanings. Across 30 task-by-model combinations it averaged AUROC 0.790, against 0.698 for P(True), 0.691 for naive token entropy and 0.687 for embedding regression. It was stable at 0.78–0.81 across model families and sizes (Farquhar et al. 2024). The same paper states the limit plainly: the method does not catch outputs that are "consistently wrong" because they were learned from training data. Its AUROC is a lower bound because some labelled errors are not confabulations at all. Verbalized confidence is the cheapest signal and the weakest. Elicited without training, it is overconfident, and failure-prediction AUROC sits at 0.522–0.605 whether the method is white-box or black-box, with every method struggling on professional-knowledge tasks (Xiong et al. 2024). Probes recover most of the sampling signal for the cost of one forward pass. Semantic-entropy probes predict semantic entropy itself at AUROC 0.70–0.95. On unseen tasks they beat probes trained on accuracy labels by +2.2 to +10.5 AUROC points, although full sampling stays more accurate (Kossen et al. 2024). A Bayesian estimator reaches semantic entropy's AUROC with 53% of the samples (Ciosek et al. 2025).

Signal family	Extra cost per query	Typical discrimination	Blind spot
Zero-shot verbalized confidence	none	AUROC 0.52–0.70	systematic overconfidence; rises with reasoning on recall tasks
Token log-prob, P(True)	~1 pass	~0.69–0.70	multi-answer questions; length bias
Hidden-state probe (SEP, exact-answer, entity)	~1 pass after one-off training	0.73–0.95 in-domain	cross-dataset transfer; "consistently incorrect" errors (0.59–0.67)
Semantic entropy / kernel entropy / SelfCheckGPT	5–20 generations + pairwise NLI	~0.79–0.80	any error the model repeats across samples
Introspective re-scoring of a reasoning trace	a second full long-CoT call	mixed	can make some models *more* overconfident

This ordering is drawn from studies that used different models and datasets, so it shows tendencies, not a meta-analysis. The more useful observation is structural: every training-free method measures instability, not error. Orgad and colleagues make this concrete. Resampling 30 answers per TriviaQA question, they sorted errors into types. Errors where the model gives "many different answers" or refuses were easy to predict from hidden states (AUC 0.85–0.90). The "consistently incorrect" type, the model's confident errors, was the hardest at AUC 0.59–0.67 (Orgad et al.). Misconceptions sharpen the point. On TruthfulQA, the best GPT-3 model was truthful on 58% of questions against 94% for humans, and larger models were *less* truthful, because they reproduce human falsehoods more fluently (Lin et al.).

The metrics hide this too. When accuracy is low, as on SimpleQA (roughly 5–30%), ECE mostly measures how *high* confidence is. Two models can have nearly identical AUROC (0.622 vs 0.613) and very different ECE (0.625 vs 0.719) (Zeng et al. 2025). A model can also minimize ECE by predicting the base rate on every item. Binned maximum calibration error *falls* with reasoning depth only because the bins fill up (Mei et al. 2025). Risk–coverage summaries (AURC, AURAC, prediction-rejection ratio) report what a selective system actually delivers (Geifman & El-Yaniv 2017; LM-Polygraph, TACL 2025). Recent reasoning-model papers almost never report Brier score or selective accuracy at fixed coverage. A reliability report should carry all three views: discrimination (AUROC/AURC), a proper score (Brier or log loss), and calibration (ECE).

Reasoning models: better calibrated where thinking helps, worse where it can't

The literature appears to conflict on whether reasoning improves calibration. Read by task type, it does not. Yoon and colleagues find that reasoning models express confidence better than their non-reasoning counterparts in 33 of 36 settings, with calibration improving as the chain unfolds (Yoon et al. 2025). Mei and colleagues find the opposite on recall-bound questions. On GPQA, more reasoning effort raised accuracy by 10.8 points and lowered ECE by 5.5. On SimpleQA, where accuracy saturates, o3-mini's ECE *rose* about 7 points from low to high effort. On questions every effort level gets wrong, confidence rose with depth, and verbalized confidence was often above 85% on incorrect answers (Mei et al. 2025). Gemini 2.5 Flash's overconfidence on IPCC confidence-level classification rose from +6% at zero thinking tokens to +21.3% at 384, while accuracy peaked at 64–192 tokens. Adding search lifted Gemini 2.5 Pro from 48.7% to 89.3% accuracy (arXiv 2508.15050). How the reasoning was trained matters as well. Large-scale RL improved calibration: DeepSeek-R1's SimpleQA ECE was 0.324 against V3's 0.515. SFT distillation into 14B and 32B models made it worse (0.719 vs 0.625 for the matched instruct model). Every reasoning variant almost stopped saying "I don't know": 2,492 abstentions for Qwen2.5-32B-Instruct against 76 and 63 for its SFT- and RL-reasoning versions (Zeng et al. 2025). The consistent reading is this. Extra thinking helps calibration when it can raise accuracy. When the answer depends on recall the model lacks, extra thinking mostly generates justification, so confidence climbs while accuracy stays flat.

E1 is a small instance of the same regime. Log-prob confidence on llama3.2 3B and llama3.1 8B predicted correctness at AUROC 0.76–0.82 on direct question forms but only 0.54–0.55 on negated forms. Both models answered "not wrong" to about 90% of false atomic claims, and answered confidently (E1). The external literature found no negation-specific confidence measurement. The coherence program's earlier notes do supply the likely mechanism: truth is represented in a two-dimensional subspace with a polarity-dependent axis, so signals tuned on affirmative statements fail on negations (Bürger et al. 2024). E1's negated-form failure is therefore a stable, polarity-driven error. Confidence had nothing to read from it.

The "do I know?" gate keys on familiarity, not on the answer

Mechanistic work from two independent groups converges on a single circuit motif. In Claude 3.5 Haiku, Anthropic's attribution graphs show "can't answer" features active *by default* on assistant prompts. "Known entity" and "known answer" features inhibit them. For "Michael Jordan" the inhibition fires and the model answers. For the invented "Michael Batkin" nothing inhibits the circuit, so the model declines. Hallucination is the misfire in between. Andrej Karpathy's name is familiar enough to partly suppress refusal, so the model names a paper he did not write. Inhibiting the known-answer features restores the refusal (Lindsey et al., "Biology of a Large Language Model").

Sparse autoencoders on Gemma 2 find linear "known entity" and "unknown entity" directions that causally control refusal. Steering along the unknown direction drives refusal to almost 100% on entities the model knows. Pushing the known direction on fabricated entities produces hallucinated attributes (Ferrando et al., ICLR 2025). The SAEs were trained on the *base* model yet steer the *chat* model. That suggests post-training wires a pretraining-era recognition signal into refusal rather than building new self-knowledge. A 2026 single-author preprint reproduces the effect on Gemma-4-12B. A single one-dimensional familiarity direction moves refusal on well-known entities from 0.24 to 1.00 and on unknown entities from 0.73 to 0.00 (Brzezinka 2026, abstract-level).

This asymmetry explains the shape of confident errors. Refusal is the default, and evidence of familiarity switches it off. Anything that raises familiarity therefore pushes toward answering: a famous subject with an obscure relation, or a name repeated in fine-tuning data or in context. Familiarity is a proxy for knowledge, and the two come apart exactly on "familiar subject, unknown fact" queries. Fine-tuning studies show the gate being pried open. On PaLM 2-S, fine-tuning examples that introduce facts the model does not know are fitted slowly. Once fitted, they raise hallucination linearly: each unit of Unknown examples costs about as much test accuracy ($\beta \approx -8.3$) as a unit of Known examples gains ($\beta \approx +7.3$). Relabelling those Unknown examples as "I don't know" lifted accuracy on answered questions to 61.8%, against 43.0% falling to 38.8% with the original labels. The share of questions answered dropped only from 58.7% to 55.6% (Gekhman et al. 2024). The proposed mechanism is that SFT on "entity X has new attribute Y" strengthens X's familiarity features and so opens the gate for X's *other* relations too. No one has tested it, and it bears directly on continual learning.

The signal is there; the decoding and policy layers don't use it

Hidden states carry more self-knowledge than the output does. Probes on the exact answer token reach AUC 0.83–0.95 on 7–8B models, against 0.63–0.75 for logit baselines. Probes at the end of the question, before any answer is generated, are weaker at 0.72–0.80 (Orgad et al.). KEEN reads the subject entity alone, generating nothing, and still correlates 0.58–0.68 with the model's QA accuracy on that entity (Gottesman & Geva 2024). Token-level probes flag fabricated entities during long-form generation at AUC 0.87–0.90 on Llama-3.3-70B, against 0.71 for long-form semantic entropy. They transfer from long-form to short-form QA but not the reverse (Obeso et al.). Across three open models, internal knowledge exceeds what is expressed by an average relative 40%. Some correct answers are ranked first internally yet never sampled in 1,000 tries, which caps what repeated sampling can recover (Gekhman et al. 2025, "Inside-Out"). Letting a probe pick among 30 samples recovered 30–40 accuracy points in the error types where the model knew the answer but did not produce it greedily (Orgad et al.).

There are three distinct kinds of "knows but doesn't say", and each needs its own fix. A correct answer can be represented but not selected by decoding; probe reranking fixes that. An error signal can be represented but ignored by the abstention gate; routing the probe into the abstain decision fixes that. And the model can have learned that guessing pays; changing the reward fixes that. The earlier coherence report adds a fourth diagnostic split. Truth probes track the model's *belief* rather than the truth (retraction study 2025). When a probe disagrees with the output, the failure is unfaithful reporting, such as sycophancy or prompt pressure. When the probe disagrees with the world, the failure is knowledge (Logical coherence of LLM world models).

Two cautions limit how far these internal signals can be trusted. First, they generalize poorly across datasets. Orgad and colleagues find most apparent cross-dataset transfer disappears once the logit baseline is subtracted. A 2025 study finds broad representation-based hallucination detectors near chance out of distribution, with no advantage over simple linear probes (arXiv 2509.19372, search-snippet level).

A probe used as a reliability signal must therefore be validated on held-out *datasets*, and the baseline it has to beat is logit confidence, not chance. Second, verbal self-report is a much weaker readout of the same states. Concept-injection experiments find Claude Opus 4 and 4.1 detecting and naming an injected concept only about 20% of the time at the best layer and strength, and the authors stress that "failures are the norm" (Lindsey 2025). Fine-tuned self-prediction beats cross-prediction by 13–17 points, but only on simple behaviors (Binder et al.). A COLM 2026 reality check finds input-only classifiers matching models' in-context self-predictions, which suggests generic anomaly detection rather than privileged access (Singh, Linzen & Ravfogel 2026, abstract-level). Introspective report is not yet a usable reliability channel; white-box probes are, where weights are available.

Binary rewards teach guessing; proper scores and labelled unknowns undo it

Confident error has three separable causes, and each has its own remedy. The first is in pretraining. Kalai and colleagues reduce generation to binary "is it valid?" classification and show that generative error is roughly at least twice the misclassification rate of that classifier. For arbitrary facts, a calibrated base model must hallucinate on at least about the *singleton rate*, the fraction of facts that appear exactly once in training (Kalai et al. 2025). The realistic target is therefore not zero error. It is unavoidable error converted into abstention, which ties lower hallucination to higher abstention on long-tail facts. The reduction fits E1 closely: E1's belief store is literally a set of yes/no validity judgments, and it is the classifier error that the generative error inherits.

The second cause is grading. Nine of the ten popular leaderboards Kalai's team examined use binary grading with no credit for "I don't know". Under that rule a guess always has higher expected score than abstaining (Kalai et al. 2025). The effect shows up in models trained against such scores. In OpenAI's o3 system card, o3 and o4-mini hallucinated on 0.33 and 0.48 of PersonQA items, against 0.16 for o1, and OpenAI attributed this partly to o3 "making more claims" (o3/o4-mini system card). Across 20 frontier models, reasoning fine-tuning lowers abstention by 24% on average. Scale helps little, and a crafted system prompt "does not resolve models' fundamental inability to reason about uncertainty" (AbstentionBench). Math reinforcement fine-tuning cuts refusal on unanswerable problems by more than 80% (Song et al. 2025; the >80% figure via a secondary review). On Qwen2.5-7B, RL on verifiable rewards improves in-domain ECE but *worsens* out-of-domain ECE, from 0.40 to 0.46 (Damani et al., RLCR).

The third cause is the preference signal. PPO post-training moved GPT-4's MMLU calibration error from 0.007 to 0.074 (GPT-4 Technical Report; values not re-verified in this round). Reward models score responses higher when they state high confidence, independent of quality (Leng et al. 2024). Human and preference-model judgments sometimes favour convincing sycophantic answers over correct ones (Sharma et al. 2023).

The remedies that work share one move: they give the abstain or low-confidence target to the model's *own* measured unknowns.

Method	What it changes	Model scale	Headline result
R-Tuning (Zhang et al.)	SFT: "I am unsure" appended where the model's answer misses gold	3–13B	refusal transfers out of domain as a "meta-skill"

Method	What it changes	Model scale	Headline result
IDK relabelling (Gekhman et al.)	SFT targets for unknown facts → "I don't know"	PaLM 2-S	answered-question accuracy 43.0% → 61.8%, coverage −3 pts
Unanswerable-data mixing (Song et al.)	10% synthetic unanswerable math in RFT	7B	unanswerable refusal 0.01 → 0.81; SelfAware 0.01 → 0.94; AMC23 up to −0.10
RLCR (Damani et al.)	reward = correctness − Brier(verbalized confidence)	7B	in-domain ECE 0.37 → 0.03 at 62.1% vs 63.0% accuracy; OOD ECE 0.46 → 0.21
Behaviorally calibrated RL (Wu et al., ICLR 2026)	Brier or log score over a prior on risk threshold t	4B	BeyondAIME confidence AUROC 0.715 → 0.902, smECE 0.632 → 0.177
TruthRL (Wei et al.)	GRPO with +1 / 0 / −1 ternary reward	not verified	−28.9% hallucination vs vanilla RL (via secondary summary)

Two independent results show what does *not* work: a discrete abstain action trained against a fixed penalty. ByteDance's team first put an explicit risk threshold t in the prompt, with reward −t/(1−t) for a wrong answer. The model over-refused even at t=0, ending below baseline accuracy, and its refusal rate barely responded to t (Wu et al.). Che and colleagues show that GRPO's group-advantage normalization rescales any designed wrong-answer penalty back to an effective 1, and that discrete abstention can collapse coverage within about 10 steps (Che et al. 2026, single preprint). TruthRL's −1 penalty worked, but −1 is the value normalization favours anyway. The robust design is a scored confidence trained with a proper rule. Its threshold k is chosen at deployment (answer iff $p > k/(1+k)$), so one trained model traces the whole risk–coverage frontier.

Calibration training lowers *confident* errors mostly by lowering confidence, not by adding correct answers. The behaviorally calibrated 4B model reached only 5.8% accuracy on SimpleQA, and its smECE there (0.459) stayed far from GPT-5's 0.098 (Wu et al.). Knowledge stays the binding constraint. The frontier illustration makes the same point. On SimpleQA without browsing, gpt-5-thinking-mini scored 22% accuracy with 26% error, against o4-mini's 24% accuracy and 75% error. About two points of accuracy bought roughly a 49-point cut in error (OpenAI, "Why language models hallucinate"; figures as tabulated by Wizwand). An abstention rate for gpt-5-thinking-mini quoted in one research note could not be verified against the source, so it is omitted here. Benchmarks that separate abstention expose the same gap: at launch, Gemini 3 Pro led AA-Omniscience on accuracy with an 88% hallucination rate on non-correct items, while Claude 4.1 Opus had lower accuracy and one of the lowest hallucination rates (The Decoder; Artificial Analysis).

System guards certify rates and inherit the model's blindness

Wrapping a model in guards changes *which* errors get through more than *whether* any do. The deciding question for each guard is whether it adds information the model lacks.

Conformal methods give marginal guarantees under exchangeability. Conformal factuality deletes low-scoring claims until FActScore, NaturalQuestions and MATH outputs reach 80–90% correctness with "the majority" of each output kept (Mohri & Hashimoto 2024). Coherent factuality calibrates over a deducibility graph and reaches 90% factuality while keeping at least 80% of claims. Calibrating *without* the graph structure fails to meet the guarantee (Rubin-Toles et al., ICLR 2025). Conformal abstention calibrates a hallucination-rate bound on the model's own sample agreement, and it abstains less conservatively than log-probability baselines on long answers (Yadkori et al. 2024). These certificates bound a population rate. They do not stop any particular confident error, and the failures can cluster. A marginal guarantee can hold while particular prompt types are badly under-covered (Cherian, Gibbs & Candès 2024). Standard conformal prediction under-covers under domain shift unless calibration points are reweighted toward the test prompt (DS-CP, ICML 2026). A 2026 preprint reports that multi-agent deployment, where models read peers' answers, breaks exchangeability: coverage on low-confidence items falls from 87% to 47% while the monitored average sits at a reassuring 76% (arXiv 2609.04445, unreplicated). Calibration fixes the threshold, not the ranking. A score that ranks confident errors as safe yields a valid certificate with poor coverage.

Deferral saves most where the cheap model is right and knows it. FrugalGPT matches GPT-4 at up to 98% lower cost (Chen et al. 2023). RouteLLM's savings drop from over 85% on MT-Bench to 45% on MMLU and 35% on GSM8K (RouteLLM project page), which are the knowledge- and reasoning-heavy benchmarks where small models are confidently wrong more often. Sequence-level confidence is length-biased as a deferral signal, and no single per-token quantile works across datasets (Gupta et al., ICLR 2024). The strongest deferral rules all learn a separate deferral function from outcome data rather than reading the first model's raw confidence. No cascade paper reports cost at a fixed *confident-error* rate, which is the quantity that matters here.

Uncertainty-triggered retrieval cannot fire on confident errors, and retrieval itself creates them. FLARE, Self-RAG and SKR retrieve when the model signals doubt (FLARE; Self-RAG). In a 35-method comparison, plain uncertainty estimators often beat purpose-built adaptive-retrieval pipelines (Moskvoretskii et al. 2025). Features that do not come from the LLM at all can match LLM-based adaptive retrieval at lower cost (EMNLP 2025). If the model is sure, though, no trigger fires. And in ClashEval, six top LLMs including GPT-4o adopted *incorrect* retrieved content over their own correct prior more than 60% of the time, more often when their own confidence was low (Wu, Wu & Zou 2024). So gating retrieval on confidence amplifies whichever way the model already leans: it defers to bad evidence when unsure and ignores good evidence when sure.

Verifiers split into information-adding and information-preserving. SAFE breaks long answers into facts and checks each one with multi-step search. It agreed with crowd annotators on ~72% of about 16k facts, was judged right on 76% of the disagreements, and cost more than 20× less than human annotation (Wei et al. 2024). Chain-of-Verification answers verification questions independently of the draft. It raised list-question precision from 0.17 to 0.36, but it is conservative: hallucinated entities fell from 2.95 to 0.68 per answer, and correct ones also fell, from 0.59 to 0.38 (Dhuliawala et al. 2023). Without new information, self-correction of reasoning often makes answers worse (Huang et al. 2023). Multi-agent debate adds almost nothing over majority voting, because it induces a martingale over agents' beliefs (Choi et al., NeurIPS 2025). Agreement between models is weak evidence on exactly the items where they are all fooled: across 349 LLMs, two wrong models pick the same wrong answer 42% of the time (Kim et al. 2025).

Sycophancy is a confident-error generator that per-query uncertainty cannot see. A bare "Are you sure?" makes models flip 46% of answers, with a 17% average accuracy drop. Synthetic fine-tuning cut the deterioration by 60% but did not remove it (Laban et al., FlipFlop). SycEval found sycophancy in 58.19%

of rebuttal cases. 14.66% were regressive (correct to wrong), citation-styled rebuttals were the most damaging, and the behaviour persisted 78.5% of the time once it started (Fanous et al. 2025). Citations doing the most damage means that the *form* of evidence, not its validity, drives adoption, the same mechanism as ClashEval. A 2026 study of nine reasoning models (GPT-5.1/5.2, DeepSeek-R1, Claude-4.5, Gemini-2.5-Pro, Grok-4.1 and others) under eight multi-turn attack types found that verbalized confidence barely predicts correctness: $r = 0.07$, ROC-AUC 0.57, mean confidence 93.5%. Flip rates were nearly flat across confidence terciles (9.3% / 6.7% / 6.2%), and a random-confidence control did as well as confidence-aware responses (arXiv 2602.13093). Treat this paper's exact figures with caution: it is internally inconsistent. Its tables give GPT-5.1's baseline accuracy as both 98.50% and 98.92%, and its version history reports the confidence–correctness correlation as both $r = 0.07$ and $r = -0.08$ (version summary). Its direction (confidence is nearly uninformative under pressure) agrees with FlipFlop and SycEval. Its point estimates should not be quoted as settled. The practical guard is to flag any answer that changes under pressure *without new evidence*, whatever its confidence.

OOD detection covers distributional novelty, not in-distribution confident errors. Relative Mahalanobis distance on input and output embeddings detects OOD inputs better than perplexity and supports selective generation (Ren et al., ICLR 2023). Token-level Mahalanobis across decoder layers serves as an uncertainty score that reportedly carries out of domain (Vazhentsev et al. 2025). LLM representations are anisotropic, so raw Euclidean distance separates poorly and whitening matters (LatentAudit). A popular-entity mistake or a sycophantic flip is not OOD, and density has no purchase on it.

The notes' own synthesis, not a published framework, maps error types to the guards that can catch them:

Error type	Self-signals (sampling, probes, verbalized)	Conformal / deferral	Retrieval / search verification	Independent model family	Density OOD
Known unknown (model is unstable)	catches	catches, with rate guarantee	catches	partial	rarely
Distributionally novel query	weak	coverage breaks	catches if evidence exists	partial	catches
Stable in-distribution wrong belief	misses (AUC ~0.55–0.67)	certifies it as safe	catches if evidence exists	catches only if errors are uncorrelated	misses
Pressure-induced flip	misses (AUROC ~0.57)	misses	can make worse (citation rebuttals)	partial	misses

NeuroGenesys's prior findings mostly survive, with one rule softened

The selector-abstention campaign, the expert-votes review and E1 were built on frozen MLP experts on Split-MNIST and on small Llama belief stores. The LLM literature tests their conclusions at a different scale and with a different error structure.

Prior finding	Literature verdict	Evidence
SA0: calibration holds where trained, fails elsewhere (own-task ECE 0.011 vs mixed 0.697)	Confirmed	RLVR improves in-domain but worsens OOD ECE (0.40→0.46); P(IK) generalizes only partly (Kadavath et al.); conformal coverage breaks under shift
SA2: put the penalty at a selector with an independent score (AURC 0.025 vs MSP 0.098)	Confirmed	every strong deferral rule learns a separate scorer; MaxProb abstention fails under shift while an external calibrator works (Kamath et al. 2020)
SA5: a penalty without out-of-competence data is inert; abstain fires uniformly	Confirmed, sharpened	explicit-t RL over-refuses and ignores t; GRPO normalizes penalties away. The LLM cures that work (R-Tuning, IDK relabelling, unanswerable mixing, HotpotQA-Modified for RLCR) all supply the model's own unknowns as out-of-competence examples
SA4 kill: training-side "not mine" did not beat a good external selector	Open at LLM scale	RLCR and behavioral calibration make training-side confidence strong, but no study compares it head-to-head with an external probe or density selector at matched data
SA1/SA7: density on a shared representation beats expert confidence	Confirmed for OOD, limited in-distribution	Mahalanobis beats perplexity for OOD; but many LLM confident errors are in-distribution, where Split-MNIST had none
SA8: grouped inputs nearly solve selection	Analogue confirmed	Mondrian and conditional conformal guarantees strengthen when the group is known
SA9 / expert-votes: errors are correlated, voting cannot fix them	Confirmed	debate is a martingale; 42% same-wrong-answer across 349 LLMs
Expert-votes thesis: never use the voter's own confidence	Partly contradicted	semantic entropy (0.79), conformal abstention on self-consistency, AutoMix self-verification and RLCR confidence all work. They work only on uncertain errors

Prior finding	Literature verdict	Evidence
E1: confidence ≈ random for repair (+1.2–1.3 pp); negated AUROC 0.54–0.55	Confirmed	the same band as "consistently incorrect" probe AUC (0.59–0.67) and pressure AUROC (0.57)
E1: LLM-judge contradiction recall 0.34–0.42	Consistent	models struggle to locate errors without new information (Tyen et al.); E1's judge is an information-preserving verifier
E1: structure beats self-report for deciding which belief yields	Consistent, extended	coherent factuality's guarantee requires the dependency graph; structural priors are an information source that does not come from the voter

The one rule the literature forces NeuroGenesys to soften is the expert-votes conclusion that "selection that does not rely on expert confidence is necessary". The Split-MNIST experts were deterministic two-class MLPs with no sampling diversity. Their only self-signal was a single softmax read, and Hein's theorem guarantees that read saturates off-distribution. An LLM has three self-derived channels the MLPs lacked. Disagreement across its own samples carries real information (semantic entropy AUROC 0.79 against 0.69 for token entropy). Its hidden states carry more (exact-answer probes up to 0.95). Its verbalized confidence can be *trained* into a calibrated, deployable score (RLCR, ECE 0.03 in-domain). The refined rule is narrower and stronger: self-derived signals are sufficient for errors the model is unstable about and insufficient for errors it holds steadily. E1 sits exactly on that boundary. The negated-form error was stable, so confidence weighting tied random. The rule that worked, entrenchment, is the kind of non-self signal the literature says stable errors require.

One NeuroGenesys finding has no external counterpart. E1 measured how often consistency repair converges on the false side: 65.0% (3B) and 31.2% (8B) of chains ended coherently wrong under confidence repair. No paper in the conformal, verification or consistency literature reports that quantity. The coherent-factuality result implies it matters, since claim-level guarantees fail when dependencies are ignored. The measurement remains NeuroGenesys's own.

Five problems remain unsolved

Nobody knows what fraction of a frontier model's errors are stable. Farquhar and colleagues call their AUROC a lower bound because some errors are not confabulations, and Orgad's taxonomy covers one 7B model on TriviaQA. No study estimates, for a current frontier model, the share of errors that are low-entropy and repeated. That share sets the ceiling for every self-derived method. A related gap: no study splits LLM confident errors into OOD and in-distribution, which sets the ceiling for density methods.

No guard has been measured on confident errors specifically. Cascade papers report cost at matched average quality. Retrieval-trigger papers report QA accuracy. None reports the rate of high-confidence wrong answers before and after the guard. The two-sided ClashEval failure predicts that confidence-gated retrieval leaves confident errors untouched and adds new ones on uncertain items. That prediction is untested.

Internal signals don't transfer across datasets, and nobody has measured transfer across checkpoints.

Correctness probes lose most of their apparent transfer once logit confidence is subtracted, and broad detectors approach chance OOD. For a continually trained model, the more important question is whether a probe fitted at checkpoint t still works at t+1. KEEN scores shift after fine-tuning, which is the only data point, and the question appears unstudied.

The fine-tuning-opens-the-gate mechanism is a hypothesis. Gekhman's linear hallucination growth and Ferrando's familiarity gate fit together. Continual-learning updates that teach new facts about an entity may strengthen its familiarity features and so suppress refusal on the entity's *other* relations. No mechanistic study tracks the known/unknown direction through fine-tuning on unknown facts, and the LoRA, instruction-tuned replication of Gekhman's result was not found.

Calibration under pressure is unmeasured where it matters. No study links calibrated confidence to resistance to pushback, that is, whether high-confidence correct answers flip less. The one 2026 study that addresses it directly is internally inconsistent. Calibration-trained models (RLCR, behaviorally calibrated RL) have not been tested under FlipFlop- or SycEval-style pressure, so it is unknown whether proper-score training buys any sycophancy resistance.

Conclusion

The question "can a model know what it doesn't know?" splits in two. Can it detect when it is *unsure*? Yes, and well: from samples, from hidden states, and from a confidence head trained with a proper score. Can it detect when it is *wrong but sure*? No, and not because current methods are immature. A stable wrong belief looks identical to a stable right one from inside the distribution that produced it. Detecting it requires information from outside that distribution: evidence, structure, an uncorrelated second opinion, or a density over a shared representation. This sorts the remedies. Training changes how much error is *expressed* as confidence: proper scoring rules, abstain targets on the model's own unknowns, and abstention-aware benchmarks. Only information-adding guards change how much error is *caught*. A system that wants to avoid being confidently wrong needs both, and should not mistake a calibrated confidence or a conformal certificate for the second.

For NeuroGenesys, this turns the selector programme's slogan into an operating rule. Trust self-derived uncertainty to flag what the model is unstable about. Route everything it marks safe through a signal the model did not produce. Report risk–coverage on the confident subset, not only average accuracy. It also points the continual-learning programme at a specific risk. Every fact written into weights is an "Unknown" example in Gekhman's sense. If the familiarity-gate hypothesis holds, each fact also opens the gate a little wider for questions the model cannot answer. An "I don't know"-relabelled arm is the cheap baseline that any reliability-preserving update method must beat.

Open problems and experiment directions

The six experiments designed from these directions are listed in the appendix below. The directions that follow most directly from this synthesis are listed below. Each runs on Qwen2.5 or Llama 3.x at 1.5–8B on the M2 Max unless noted.

Stable-error census. Resample 30 answers per item on TriviaQA, EntityQuestions and E1's MQuAKE beliefs. Measure the share of "consistently incorrect" errors, and compare semantic entropy, an exact-answer probe and log-prob confidence on that subset alone. This gives the missing denominator for every self-signal claim.

Familiarity-gate tracking through fine-tuning. Label EntityQuestions with SliCK, then LoRA-fine-tune

on 0/25/50/100% Unknown mixtures with an "I don't know"-relabelled arm. At each checkpoint, track a known/unknown entity direction on held-out relations of the fine-tuned entities, together with exact-answer probe drift. This tests the gate hypothesis and the probe-across-checkpoints gap together.

Proper-score versus ternary versus external selector at matched compute. Compare RLCR's Brier reward, TruthRL's +1/0/−1 reward and a frozen-feature probe or density selector on a 1.5B model, scored by the Kalai +1/0/−t/(1−t) sweep already implemented for SA2. This is the LLM-scale version of the SA4 kill. It needs one or two rented H100s for the RL arms.

Confident-error accounting for guards. Measure the high-confidence error rate before and after uncertainty-triggered retrieval, a learned cascade and an NLI-model verifier, with ClashEval-style corrupted evidence in one arm.

Pressure on calibrated models. Run FlipFlop and SycEval rebuttal protocols on base, RLCR-trained and IDK-relabelled checkpoints. Test whether calibrated confidence predicts which answers flip, and whether a probe–output disagreement flag catches the flips that confidence misses.

Appendix: the SK1–SK6 experiment plan

Six experiments were designed from this synthesis. None has results yet, and no result numbers are reported here. SK1–SK4 are pre-registered and queued; SK5 and SK6 are designed in the plan but not yet pre-registered, because they need rented GPUs. Each confirmatory family is Holm-corrected.

ID	Question	Status and hardware
SK1	Does sampling disagreement beat a single-read confidence once errors are stable or the user pushes back?	Pre-registered, queued; local CPU (Ollama, llama3.2 3B and llama3.1 8B, with qwen2.5-coder 7B for cross-family disagreement)
SK2	What does uncertainty-triggered retrieval do to the *confident*-error rate?	Pre-registered, queued; local CPU, sharing SK1's item bank
SK3	Does teaching a model a few facts about an entity open its "known entity" gate for every question about it?	Pre-registered, queued; local MPS (Qwen2.5-1.5B with LoRA)
SK4	What does a model cascade cost at a fixed confident-error rate, rather than at fixed average accuracy?	Pre-registered, queued; local CPU, offline analysis of SK1 and SK2 outputs
SK5	Can an exact-answer probe rescue consistently wrong items that sampling cannot catch?	Designed, not yet pre-registered; needs a rented GPU (7–8B model)
SK6	Brier-scored confidence against a ternary abstain reward at matched compute, and does SFT-taught abstention survive later RL?	Designed, not yet pre-registered; needs rented GPUs (Qwen2.5-1.5B GRPO)

SK1 defines a stable stratum (items whose modal answer recurs in at least 8 of 10 samples) and predicts that sampling agreement beats single-read confidence overall but is near chance on stable items, where cross-family disagreement should do better. It also preregisters the missing number this synthesis identifies: the share of all errors that are stable. A pushback arm tests whether pre-pressure sampling agreement predicts which answers flip under a citation-styled rebuttal, and whether reverting answers that changed without new evidence cuts confident errors.

SK2 predicts that a confidence-triggered retrieval policy recovers most of always-retrieve's accuracy gain while barely touching errors the model was confident about, doing worse on them than a random trigger at the same budget. A conflicting-evidence arm measures whether retrieval creates new confident errors, mostly on items the model was unsure of, and tests a cheap evidence-disagreement guard.

SK3 tests the familiarity-gate hypothesis: LoRA updates that teach a few unknown facts about an entity should lower abstention on that entity's held-out relations without making those answers more accurate, with the effect growing with facts per entity, running through a measurable familiarity direction, and shrinking when "I don't know" targets are trained on another of the entity's relations but not under a KL anchor.

SK4 costs a 3B → 8B → terminal-tier cascade at a fixed rate of unflagged errors and predicts that this criterion both raises the bill and changes which deferral rule wins, because the two models share many confident errors.

SK5 asks whether an exact-answer probe ranks the gold answer above a model's own stable wrong answer when the gold is offered as a candidate, separating decoding failures from missing knowledge.

SK6 compares a Brier-scored confidence reward with a $+1/0/-1$ abstain reward under GRPO at matched compute, tests whether penalty normalisation collapses coverage at 1.5B, and tests whether abstention taught by SFT survives later RL on verifiable rewards.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/model-self-knowledge-confident-errors