How LLMs Detect and Correct Their Own Errors: The Role of Internal Confidence Signals

Anonymous authors
Paper under double-blind review

Abstract

Large language models can detect their own errors and sometimes correct them without external feedback, but the underlying mechanisms remain unknown. We investigate this through the lens of second-order models of confidence from decision neuroscience. In a first-order system, confidence derives from the generation signal itself and is therefore maximal for the chosen response, precluding error detection. Second-order models posit a partially independent evaluative signal that can disagree with the committed response, providing the basis for error detection. Kumaran et al. (2026) showed that LLMs cache a confidence representation at a token immediately following the answer (i.e. post-answer newline: PANL)—that causally drives verbal confidence and dissociates from log-probabilities. Here we test whether this signal supports error detection—the core prediction of the second-order framework—and a hypothesized extension to self-correction. Using a verify-then-correct paradigm, we show that: (i) verbal confidence predicts error detection far beyond token log-probabilities, ruling out a first-order account; (ii) PANL activations predict error detection beyond verbal confidence itself; and (iii) PANL predicts which answers the model will revise, and critically which errors the model can correct—beyond all behavioural signals. Causal interventions establish that PANL signals rescue error detection when answer information is corrupted, and that ablating answer-adjacent activations including PANL disrupts self-correction. All findings replicate across models (Gemma 3 27B and Qwen 2.5 7B) and datasets (TriviaQA, PopQA, and MNLI). These results reveal that LLMs naturally implement a second-order confidence architecture whose internal evaluative signal encodes not only whether an answer is likely wrong but whether the model has the knowledge to fix it.

1 Introduction

The ability of large language models to detect and sometimes correct their own errors without external feedback (e.g., Huang et al. 2024; Kamoi et al. 2024) is well documented but poorly understood. Confidence ratings can be obtained through diverse means—token log-probabilities (Kadavath et al., 2022), sampling-based consistency (Geng et al., 2023; Tian et al., 2023), and verbal reports (Xiong et al., 2023; Yoon et al., 2025; Steyvers et al., 2025). In biological agents, confidence drives error monitoring, help-seeking, and strategy revision (Rabbitt, 1966; Yeung et al., 2004; Fleming & Daw, 2017). The self-correction literature, however, has studied *whether* models can improve their answers without asking a fundamental prior question: does the model's own confidence predict whether it will identify an error and whether a correction attempt will succeed?

We address this question at two levels: behaviourally, we ask whether verbal confidence predicts verification and self-correction outcomes; representationally, we ask whether the model's internal activations carry richer information than its overt confidence report. Kumaran et al. (2026) showed that LLMs cache a confidence representation at the post-answer newline token (PANL) that causally drives verbal confidence and dissociates from generation-time log-probabilities. Here we ask whether this same signal extends beyond confidence to support error detection and self-correction, drawing on the distinction between first-order and second-order models of confidence from decision neuroscience (Fleming & Daw, 2017; Kepecs et al., 2008; Kiani & Shadlen, 2009) (see Figure 1 and Appendix B).

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.

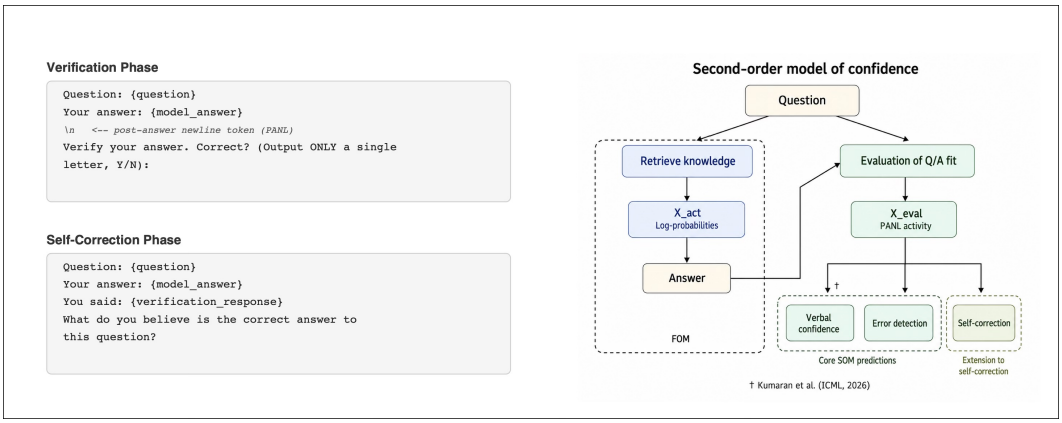

Figure 1: **Left panel**: Verification and self-correction prompt structure (see §A.2 for full details). The model's answer and verbal confidence were generated in a separate prior phase. In the verification phase, the model is shown its own answer to a TriviaQA (or PopQA/MNLI) question and asked to judge whether it is correct (Y/N), followed by a self-correction prompt. Residual stream activations were extracted during the *verification* phase at the post-answer newline token (PANL, indicated by arrow)—the first token after the model's answer following (Kumaran et al., 2026). **Right panel**: Second-order model of confidence, adapted from Fleming & Daw (2017). Left side (dashed box): the first-order model (FOM), in which a generation process produces an answer and the associated log-probabilities (X_{act}) are the only available confidence signal. Under greedy decoding, X_{act} — and therefore confidence—is by definition maximal for the chosen answer, so a purely first-order system cannot conclude it erred. Right side: the second-order model, in which the completed answer engages a qualitatively distinct evaluative process that assesses question–answer fit by attending backward over the full response—a different computation from the retrieval process that produced it (see §B). The resulting evaluative signal (X_{eval}; termed X_{conf} in the original framework), encoded at answer-adjacent token positions (PANL), is partially independent of X_{act}. A core prediction of the 2nd order framework is that X_{eval} drives verbal confidence (i.e. as shown in Kumaran et al. (ICML, 2026), and error detection. According to our extension of the 2nd order framework, X_{eval} is also involved in self correction.

In first-order models, a single signal drives both the decision and the confidence report—the decision variable X_{act} determines both which response is selected and confidence in it. Applied to LLMs, the first-order analogue is token log-probabilities: under greedy decoding, confidence is by definition maximal for the chosen answer, leaving no basis for error detection. In second-order models, confidence arises from a partially independent variable X_{eval} that performs a qualitatively distinct evaluation of the response relating to question-answer fit. Because X_{eval} is not yoked to X_{act}, it can disagree with the committed response—providing the foundation for error detection (Fleming & Daw, 2017)(see §B).

The second-order framework generates specific predictions that we test in this study. Two follow directly from the core architecture (see Figure 1): if the model maintains an evaluative signal at PANL that is partially independent of the generation signal, then (i) verbal confidence should predict error detection beyond token log-probabilities, since the evaluative signal carries information that the generation signal does not; and (ii) PANL activations should predict error detection beyond verbal confidence itself, since the model's internal evaluation is only partially externalised in its confidence report. Beyond these core predictions, we hypothesize an extension of the second-order framework to self-correction: if the evaluative process has independent access to the model's knowledge—rather than merely monitoring the committed response—then (iii) PANL activations should predict which answers the model should revise, and importantly which revisions will be correct, above and beyond behavioural confidence signals (see Figure 1). We test these predictions in non-reasoning models—without chain-of-thought there is no reasoning trace to re-read, so error detection can only arise from a signal that directly evaluates question–answer fit, distinct from the generation process itself: a pure test of the first-order/second-order distinction.

Our main contributions are:

1. **Behavioural second-order signature (core prediction).** Verbal confidence predicts verification responses beyond token log-probabilities, ruling out a first-order account—though it predicts only error detection, not correctability.
2. **PANL plays a causal role in error detection (core prediction).** PANL predicts which answers the model will flag as errors, beyond behavioural signals. Patching shows PANL is causally sufficient to rescue error detection; joint ablation shows the signal is shared redundantly with the last answer token (LAT), PANL isolating the evaluative component from answer content.
3. **PANL predicts correctability (extension).** PANL activations predict which revisions will succeed—where behavioural signals fail—and causal noising implicates PANL, jointly with LAT, in the corrective computation.
4. **Generality.** All key findings replicate across Gemma 3 27B and Qwen 2.5 7B and across TriviaQA, PopQA, and MNLI; error-detection probes transfer near-perfectly between factual QA datasets, indicating a domain-general evaluative direction.
5. **Second-order architecture.** Together, these results provide the first evidence that LLMs spontaneously implement a second-order confidence architecture (Fleming & Daw, 2017)—an evaluative process partially independent of generation that can disagree with the committed response—and support our hypothesized extension to self-correction.
6. **Practical relevance.** A PANL probe enables selective self-correction: routing the top half of trials by probe score matches unconditional self-correction at half the inference cost.

2 Related Work

We briefly describe the most closely related work (see Appendix B for a detailed review). LLMs have variable capacities to detect their own errors and self-correct (termed intrinsic self-correction) (Huang et al., 2024; Kamoi et al., 2024; Zhang et al., 2025; Madaan et al., 2023; Liu et al., 2024; Chen et al., 2025; Weng et al., 2023). Li et al. (2024) prompt the model to maintain its answer if confident and update otherwise, but do not examine whether confidence signals predict error detection or downstream correction success. Bertolazzi et al. (2025) use circuit analysis to study arithmetic error detection in small LLMs (1.5B–3B), identifying "consistency heads" that perform surface-level digit matching—locating circuit components rather than characterising the representations involved, and without studying confidence.

A growing body of work shows that LLMs encode output quality in internal activations, often diverging from surface-level confidence (Azaria & Mitchell, 2023; Liu et al., 2025; Burns et al., 2022; Li et al., 2023; Bürger et al., 2024; Orgad et al., 2024; Lu et al., 2025). A complementary literature studies confidence *calibration*—how well elicited confidence tracks first-pass correctness (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2023; Geng et al., 2023; Steyvers et al., 2025). We instead ask whether confidence signals predict error detection, revision, and correction success—targets a well-calibrated first-pass score need not capture—and our framework explains the recurring dissociation between verbalised and logit-based confidence (Xiong et al., 2023) as two distinct signals (X_{eval} vs X_{act}) rather than noisy readouts of one.

Kumaran et al. (2026) showed that verbal confidence is not computed on-the-fly: the model caches a confidence representation at the post-answer newline token (PANL), later read out as verbal confidence. Under causal attention, PANL attends to all question and answer tokens but no subsequent prompt tokens, making it a backward-looking summary of the completed response. PANL activations at middle layers causally drive verbal confidence and explain variance *beyond* answer log-probabilities—a richer evaluation of question–answer fit than generation fluency. The present paper asks whether this same signal supports error detection and self-correction.

3 Experiments

We study Gemma 3 27B (Team et al., 2025) as our primary model on the TriviaQA factual knowledge dataset (Joshi et al., 2017) ($n = 7{,}227$ questions), with cross-model replication on Qwen 2.5 7B ($n = 3{,}500$) and cross-task replication on PopQA(Mallen et al., 2023) and MNLI (Williams et al., 2018). All answers were generated with greedy decoding (temperature $= 0$), ensuring that the initial answer (A1) represents the model's argmax of its token distribution. Any improvement from initial

answer (A1) to the second answer (A2) therefore cannot arise from stochastic resampling. Full model and dataset details are in §A.1.

In the simple verify-then-self-correct paradigm we employ, the model generates an answer and reports its confidence (Phase 0), is then shown its own answer and asked to verify it ("Y"/"N"; Phase 1 (verification)), and finally produces a second answer (Phase 2 (self correction); see §A.2; Figure 1). We extract residual stream activations at PANL during the verification phase and use linear probes to predict verification behaviour and second-attempt correctness. A1 and A2 correctness were scored with GPT-4o, validated against both exact string matching and human judgment (see §A.2). We first characterise the predictive value of behavioural signals and then turn to the activation analyses to ask whether PANL representations explain variance that behavioural signals leave unaccounted for.

3.1 Basic Behavioral Results

Gemma 3 27B scored 75.5% on the TriviaQA dataset at its first attempt (A1: $n = 7{,}227$ questions). Verbal confidence and the log-probabilities of the answer tokens yielded an AUROC of 0.74 and 0.76 for predicting A1 correctness, respectively; expected calibration error (ECE) was 0.13 and 0.20, respectively. Self-correction yielded a modest but highly significant improvement in accuracy, from 75.5% at A1 to 79.2% at its second attempt (A2; $\Delta = +3.7$ percentage points; McNemar's $\chi^2 = 202.3$, $p < .001$).

Prediction 1: Gemma shows a reliable error detection capacity Figure 2 shows verification responses as a function of A1 correctness. The model displays robust error-detection sensitivity ($d' = 1.67$) with a strongly liberal criterion ($c = -1.34$), reflecting a pronounced bias to respond Yes. A foil experiment confirms this bias is not a fixed property: when the model verifies answers it did not generate, error detection ability scales with foil plausibility ($d' = 2.57$ for hard foils to 5.08 for unrelated foils) and the Yes-bias diminishes accordingly ($c = -0.78$ to -0.14; see §A.2.1, §F.2, and Figure A4).

Prediction 2: Verbal confidence carries predictive information beyond log-probabilities for predicting verification behavior. (Core prediction of second-order framework) A second-order account predicts that verbal confidence, because it is driven by an evaluative signal partially independent of the generation process, should predict the verification response beyond what token log-probabilities alone can explain. Note that although a reliable verbal confidence signal should by definition predict binary correctness of the initial response (as it does, AUROC = 0.74 for A1 correctness reported above), it need not necessarily predict verification response.

We tested the role of verbal confidence in predicting the verification response (V=Y vs V=N) with nested logistic regressions. A model containing only mean answer log-probability yielded AUROC $= .668$ (see §A.5 for method); adding verbal confidence produced a dramatic improvement (AUROC $= .832$; LR $\chi^2 = 839.4$, $p < 10^{-184}$; with standardised coefficients confirming that confidence dominates, $\beta_{\text{conf}} = +0.51$, $\beta_{\text{logprob}} = +0.10$). The pattern is even clearer within incorrect trials, where log-probabilities are entirely uninformative about the verification response (AUROC $= .481$) while verbal confidence alone achieves AUROC $= .737$ (LR $\chi^2 = 164.6$, $p < 10^{-37}$); log-probabilities add nothing once confidence is included ($\chi^2 = 0.37$, $p = .54$). Verbal confidence thus carries substantial evaluative information about the model's own answer that generation-time log-probabilities do not capture (see Figure 2D). This dissociation is robust to enriching the log-probability baseline with additional generation-time statistics (minimum and first-token log-probability, answer length; §F.1), which add only marginal predictive value. These findings are consistent with verbal confidence reflecting an evaluative process at least partially independent of generation fluency.

Prediction 3: PANL activations predict verification behavior better than verbal confidence (Core prediction). To test whether PANL representations carry evaluative information beyond what behavioural signals capture, we trained linear probes on residual stream activations at the PANL position (see §A.4) and extracted a cross-validated scalar *probe score*. Note that probes are trained on the model's verification response, not on ground truth: they predict the model's own evaluative judgment—including its false alarms (i.e. errors afforded a Yes response)—rather than serving as an oracle for answer correctness; the fidelity of this judgment to ground truth is the model's error-detection sensitivity itself ($d' = 1.67$). Figure 3A shows that PANL activation at mid-to-upper

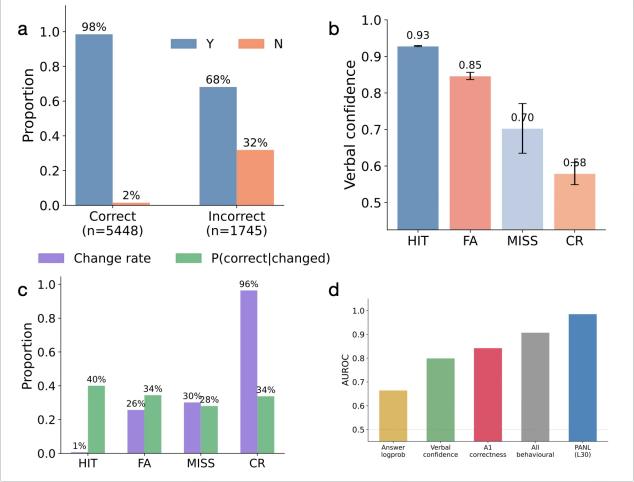

Figure 2: Gemma 3 27B behavioural results on TriviaQA ($n = 7{,}223$). **a**: Verification responses by A1 correctness. The model displays robust error detection ($d' = 1.67$) with a strong Y-bias ($c = -1.34$): it confirms 98% of correct answers but still endorses 68% of incorrect ones. **b**: Mean verbal confidence ($\pm$95% CI) across SDT cells (e.g. Hit = A1-correct given a Y response at verification (i.e. V=Y). FA: A1-incorrect given V=Y). **c**: Answer change rate (purple) (i.e. proportion of answers changed between A1 and A2) – and correction success conditional on changing (green) by SDT cell. Verification gates whether the model revises, but correction success is roughly constant ($\sim$28–40%) regardless of cell. **d**: AUROC for predicting the verification response from answer token log-probabilities, verbal confidence, A1 correctness, their combination and PANL activation at layer 30. Verbal confidence substantially outperforms log-probabilities, consistent with a second-order evaluative signal that is partially independent of generation fluency. PANL outperforms best behavioral baseline.

layers substantially exceeds the combined behavioural baseline (log-probabilities, verbal confidence, A1 correctness; AUROC = .986 vs .908; LR $\chi^2 = 1100.0$, $p < .001$; Table A1). The advantage is far larger within incorrect trials (Figure 3B)—where the verification decision is most consequential and behavioural signals are weakest—with PANL achieving AUROC = .958 against a baseline of .715 (Figure 4A, Table A2; Figure 3D shows the probe separating correct rejections from false alarms). This supports P3: the model's evaluative representation contains more information about its own accuracy than it communicates in its overt confidence report—verbal confidence is a partial, lossy readout of a richer internal signal.

Beyond the binary verification decision, PANL activations predict the *strength* of that decision: the probe score tracks the verification logprob difference ($r = .91$; Ridge prediction exceeds the behavioural baseline $r = .715$ at mid-to-upper layers; Figure A2D), while correlating only moderately with verbal confidence ($r = .60$) and barely with answer log-probability ($r = .21$; .02 within incorrect trials)—the profile of a pre-computed evaluative signal that drives verification, is partially externalised in confidence, and is decoupled from generation fluency.

The predictive advantage of PANL survives controlling for a question-level index of item answerability, indexed by P(IK)(see §A.3)—the proportion of correct responses across 20 samples (Kadavath et al., 2022). PANL adds above a baseline including P(IK) both overall (AUROC = .978 vs .909 on the subset with P(IK) available; LR $\chi^2 = 154.3$) and within incorrect trials (.936 vs .759; $\chi^2 = 33.7$; both $p < 10^{-8}$), and the verification and P(IK) probe directions are near-orthogonal at layer 30 (cosine = $+0.007$; §F.3). The evaluative signal at PANL is thus an *instance-specific* judgment of this question–answer pair, not a reflection of *question-level* difficulty.

Figure 3A also shows that the last prompt token has comparable predictive power to PANL, as do downstream positions (PANL+1, offset 6–18; Figure A2A)—the former unsurprisingly, since it directly precedes the Y/N decision; the latter carry decodable but not causally relevant information (§F.4). The last answer token (LAT) also carries substantial predictive information—we focus on PANL because LAT is part of the answer representation itself: any intervention at LAT disrupts

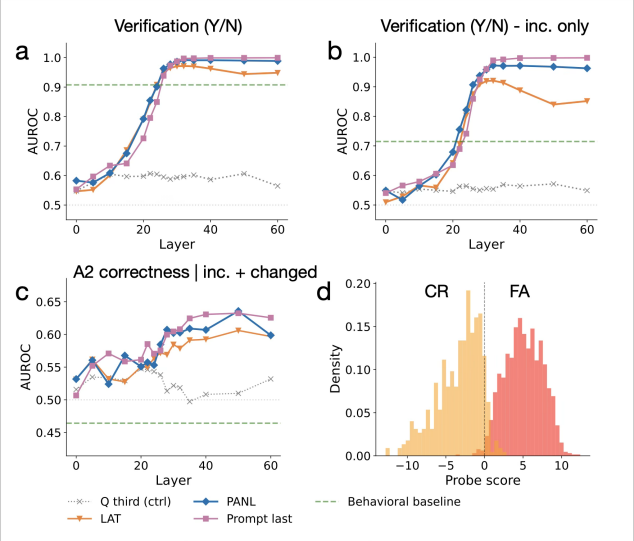

Figure 3: Linear probing results across layers for three targets: **A**: verification response (all trials), **B**: verification response (incorrect trials only), **C**: A2 correctness (incorrect trials that changed answer). Probes are L_2-regularised logistic regression ($C = 0.001$, 5-fold CV) on residual stream activations at four token positions: question third token (control), last answer token (LAT), post-answer newline (PANL), and prompt last token. The dashed green line indicates the best behavioural baseline (see Table A1). PANL exceeds the behavioural baseline across all three targets at mid-to-upper layers, with the largest margin for A2 correctness among changed incorrect trials (C), where behavioural predictors are at chance. The control position shows no improvement for any target. **D**: Distribution of PANL L30 probe scores (trained on verification Y/N) within incorrect trials, split by verification response. The probe cleanly separates correct rejections (CR; model detected its error) from false alarms (FA; model missed its error), with minimal overlap near the decision boundary (AUROC = .892). This signal is encoded at PANL during the answer phase, before the verification instruction is presented.

both answer content and any evaluative signal computed over it, confounding the two in causal experiments.

The evaluative signal predicts answer revision. The model changes its answer far more often when it reports an error (V=N) than when it endorses it (V=Y): within incorrect trials, 97% of correct rejections led to an answer change compared with 27% of false alarms (Figure 2C). Phase-0 behavioral measures predict revision only modestly (answer log-probability AUROC = .508; verbal confidence = .657; Table A3). The verification logprob difference—a continuous measure of the strength of the verification decision—is far stronger (AUROC = .907), as expected given that revision is gated by the verification response. PANL improves even on the combined behavioural baseline including this signal (AUROC = .921 vs .901; LR $\chi^2 = 237.7$, $p < .001$; Table A1): since the probe score pre-computes the verification decision itself ($r = .91$), its advantage over the verification logprob difference indicates that the internal representation carries finer-grained information about revision than its own binary readout conveys.

Prediction 4: PANL predicts which errors the model can correct (Extension) The preceding analyses show that the evaluative signal at PANL predicts both whether the model detects an error and whether it revises its answer. But the most consequential question for self-correction is whether the model can predict which revisions will *succeed*–our extension of the 2nd order framework. We focus on incorrect trials that changed their answer ($n = 856$)—the only subset where correctability prediction is non-trivial, since unchanged correct answers dominate A2 correctness elsewhere. Conditional on changing its answer, the model produces a correct A2 approximately one-third of the time regardless of SDT cell (∼29–34%; Figure 2C). The combined behavioural baseline falls below

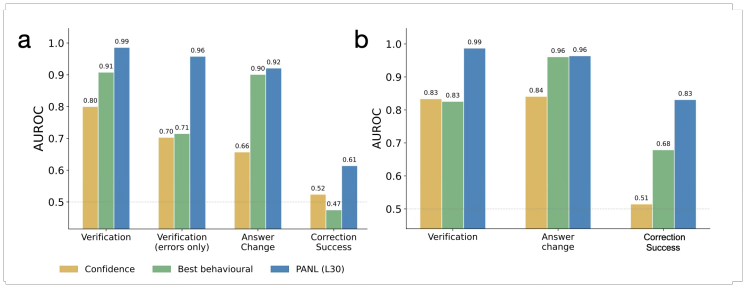

Figure 4: **a**) Main experiment: AUROC for predicting verification response (all trials), verification response (incorrect trials only), answer change (incorrect trials), and A2 correctness (incorrect trials that changed answer). *Verbal confidence*: the model's stated confidence from Phase 0. *Best behavioural baseline*: logistic regression over available behavioural scalars—answer log-probability, verbal confidence, and A1 correctness for verification; answer log-probability, verbal confidence, and verification log-probability difference for answer change and A2 correctness. *PANL probe*: linear probe on post-answer newline activations at layer 30. The PANL probe matches or exceeds the best behavioural baseline across all targets, and is the only predictor substantially above chance for A2 correctness among changed trials (AUROC $= .614$ vs. $.475$ for the behavioural baseline; see Table A1 for summary; Tables A2–A4 for individual predictor breakdowns). **b**) Analogous results for the foil experiment, averaged across hard, easy, and unrelated foil conditions. Verification within incorrect trials is not shown because the candidate answer is always incorrect by construction. See Table A5 for per-condition results.

chance (AUROC $= .475$), confirming that no combination of the model's overt signals distinguishes correctable from uncorrectable errors once the decision to revise has been made.

The PANL probe score, however, predicts correction success with AUROC $= .614$ (LR $\chi^2 = 9.5$, $p < .01$; Table A1; Figure 4A), visible across mid-to-upper layers (Figure 3C). This performance is modest in absolute terms—though four further settings bolster the finding.

First, in the foil experiment a PANL probe predicts correction success with AUROC .813–.859 across foil types—substantially higher than for own-answer trials (Figure 4B; likely reflecting the larger, more varied pool of answer-change trials in foil conditions)—while behavioural predictors remain at or near chance (Table A5). Second, on MNLI (Appendix G.2)—a cross-task test outside factual QA—PANL predicts correction success with AUROC $= .862$ against a behavioural baseline at chance (.512). Third, the PopQA dataset (Section G.4) provides a considerably more demanding test of correctability prediction than TriviaQA: base accuracy is far lower (29.3% vs 75.5%), yielding a much larger and more heterogeneous pool of incorrect trials that changed answer ($n = 3{,}688$) on which the probe can be trained and evaluated. On this dataset PANL predicts correction success with AUROC $= .785$ against a behavioural baseline of .671 ($\Delta = +.114$). Fourth, the causal noising experiments (see below) show that the corrective computation itself depends on answer-adjacent activations, with PANL contributing jointly with LAT in both models—the correctability signal is not merely decodable but causally implicated in correction. Across these settings PANL predicts correctability with mean AUROC $= .735$ (range .614–.862) while the behavioural baseline sits at or near chance except on PopQA, where PANL still adds significantly beyond it (Table A6). These findings support P4: PANL encodes information about whether the model has the knowledge to produce a better answer, captured by no behavioural signal, including the verification decision itself. Within the second-order framework, this is the signature of an evaluative process with access to the model's knowledge structure: the same process that assesses question–answer fit also provides evidence about what the correct answer should be.

These results survive controls for knowledge-level and surface confounds: on TriviaQA, PANL adds beyond baselines augmented with P(IK) (answer change LR $\chi^2 = 101.7$; A2 correctness $\chi^2 = 99.0$; both $p < 10^{-23}$) and surface text features add nothing over PANL ($p = 1.0$; §A.4). On PopQA, a maximal first-order baseline combining behavioural predictors with P(IK) and an entity-popularity proxy reaches AUROC $= .77$, approaching PANL alone (.785)—knowledge explains much of correctability where it is well measured—yet PANL still adds significantly above it ($\chi^2 = 104.3$,

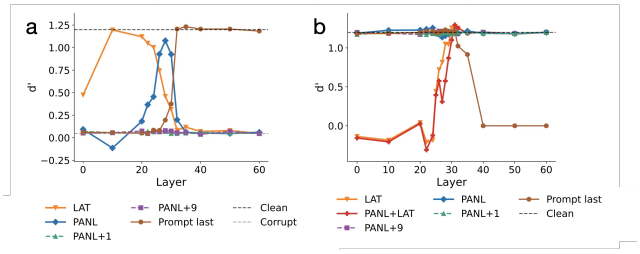

Figure 5: Gemma 3 27B causal experiments on TriviaQA ($n = 1{,}000$). **a**: Activation patching. Clean $d' = 1.17$; corrupted $d' = 0.09$ (grey dashed). LAT (orange) rescues at early-to-mid layers (peak $d' = 1.23$ at L15, $\sim$100% recovery); PANL (blue) rescues at mid layers (peak $d' = 0.89$ at L30, $\sim$74% recovery); prompt last token (brown) rescues at later layers (peak $d' = 1.25$ at L35). PANL+1 and PANL offset 9 show no rescue despite carrying decodable information. **b**: Activation noising through mean ablation. Ablating LAT alone (orange) severely disrupts error detection at early-to-mid layers ($d' = -0.21$ at L22), recovering by L30. Ablating PANL alone (blue) has no effect. Jointly ablating LAT and PANL (red) produces additional deficits at mid layers where LAT is recovering (L27: $d' = 0.31$ vs 0.82 for LAT alone), consistent with representational redundancy. Ablating the prompt last token (brown) disrupts at later layers ($d' = 0.0$ at L40–60).

$p < 10^{-24}$; §G.4; Figure A7f). PANL thus encodes a trial-specific evaluative signal—whether *this* answer fits *this* question—not answerability, entity popularity, or surface text (also see §F.3).

Summary of behavioral and linear probing analyses. Together, these results confirm the core predictions of the second-order framework and support our hypothesized extension. Regarding core predictions: the model reliably detects its own errors (P1); verbal confidence carries evaluative information beyond log-probabilities (P2); and PANL activations predict verification beyond all behavioural signals, particularly within incorrect trials where the gap is largest (P3). On the extension: PANL predicts which errors the model can correct, where behavioural signals fail (P4)—indicating that the evaluative process has access to the model's knowledge, not merely to the quality of the committed response. The evaluative signal at PANL is thus richer than what the model externalises in its overt confidence report, or the strength of the verification decision—consistent with PANL encoding an evaluation of question–answer fit that indexes not only whether the answer is likely wrong, but whether the model has the knowledge to fix it (Figure 1). We next ask whether this signal plays a causal role in error detection and self correction.

4 Causal Interventions

To establish whether the evaluative information at PANL plays a causal role in error detection, we performed activation patching and noising experiments on Gemma 3 27B ($n = 1{,}000$ TriviaQA trials; see §A.6 and §A.7 for methods; see §G.9 for Qwen 2.5 7B results).

PANL is causally sufficient for error detection. We corrupted the verification prompt by replacing all answer token embeddings with position-specific means, abolishing error detection (d': $1.17 \rightarrow 0.09$). Restoring clean activations at a single position and layer reveals three causally relevant positions with complementary temporal profiles (Figure 5A; see Figure A3A for plot of verification log-probability difference (Y-N)): LAT rescues at early-to-mid layers ($\sim$100% recovery at L15), PANL at mid layers ($\sim$74% recovery at L30), and the prompt last token at later layers ($\sim$107% recovery at L35). Positions where linear probes decode verification-predictive information (PANL+1, PANL offset 9) show no rescue at any layer, confirming that decodable information is not always causally expressed in behaviour. The temporal dissociation between LAT and PANL is consistent with an evaluative computation that begins at the answer tokens and is consolidated at PANL in mid-to-upper layers—the same layers where probing performance peaks.

PANL is not necessary for error detection, but carries the evaluative signal redundantly with LAT. Mean ablation—replacing a position's activation with its layer-specific mean from a balanced

calibration set—reveals a complementary pattern (Figure 5B; see Figure A3B for plot of verification log-probability difference (Y-N)). Ablating PANL alone has no effect at any layer ($|\Delta d'| \leq 0.062$), while ablating LAT severely disrupts error detection at early-to-mid layers ($d' = -0.21$ at L22), recovering by L30. However, jointly ablating LAT and PANL produces deficits beyond LAT alone at mid layers (e.g., $d' = 0.31$ vs 0.82 at L27), precisely where PANL's probing performance peaks. This additional deficit is only visible when LAT is also disrupted—consistent with the two positions carrying the evaluative signal redundantly (see §F.4 for details, and why LAT confounds answer and evaluative representation).

Causal evidence extends to self-correction. Applying mean-ablation noising to a one-step correction prompt (see §A.8 and §F.5 for methods and full results; Figure A5), we find that the corrective computation depends on the same two positions as verification: ablating LAT substantially reduces the probability of producing the correct answer in both models, and jointly ablating LAT and PANL produces deficits beyond LAT alone at mid layers—the same redundancy signature observed for error detection. Notably, LAT ablation *suppresses* correction rather than facilitating a switch away from A_1, indicating that LAT carries evaluative information over the committed answer rather than a literal token copy. Note that patching is not a principled intervention here—corrupting the answer span removes the very answer self-correction must operate on—so noising serves as the primary causal test.

Cross-model, cross-task generalisation and practical relevance Qwen 2.5 7B replicates all key findings (see §G.8 and §G.9 for full results; Figures A9 and A10): robust error detection ($d' = 1.63$), PANL predicting verification, answer change, and correction success beyond behavioural baselines (AUROC $= .774$ for A2 correctness within incorrect trials, baseline $.568$), and the same causal architecture—PANL sufficient, LAT necessary, both redundant at mid layers. We also demonstrate that the findings generalize to a different more challenging factual QA dataset - PopQA (see §G.3 and §G.4 and Figure A7). To test whether the key findings—in particular PANL's ability to predict correction success—are specific to factual QA, we replicated the paradigm on MNLI (three-way entailment classification), restricting analyses to neutral trials where verification varies meaningfully (§G.1, §G.2; Figure A6). Despite much weaker behavioural signals, PANL replicates all probing results, including correction success within incorrect changed trials (AUROC $= .862$, baseline at chance).

Cross-dataset probe transfer separates domain from format: verification probes transfer near-perfectly between TriviaQA and PopQA (AUROC $= .982/.988$), which share the factual QA domain but differ in question structure, while transfer to MNLI—a different domain and answer format—is substantially weaker. Correctability transfer follows the same ordering at lower levels, with low cosine similarity between correctability directions across both dataset pairs ($.071$ vs $.397$ for TriviaQA–PopQA correctability vs verification): a correctability signal exists universally at PANL, but its direction is more task-specific than the verification direction (Table A8; §G.5).

Evidence also suggests the PANL signal is practically deployable: a probe trained to predict when self-correction will help routes only half of trials to the correction prompt while matching the accuracy of correcting every trial—halving correction-time inference cost (AUROC $= 0.90$ TriviaQA, 0.86 PopQA; Table A9; see §G.6 for details).

5 Conclusion

Kumaran et al. (2026) showed that LLMs cache an evaluative representation at PANL that drives verbal confidence and dissociates from log-probabilities. We show that this signal supports error detection—confirming the core prediction of the second-order framework from decision neuroscience (Fleming & Daw, 2017). Going beyond the core framework, we find that the same signal predicts which errors the model can *correct*, beyond all behavioral signals—supporting our hypothesized extension in which the evaluative process has independent access to the model's knowledge. Finally, we speculate that this automatic error signal may be what reasoning models have learned to act on: redeployed at intermediate commitment points within a reasoning trace to trigger backtracking (Ward et al., 2025; Gandhi et al., 2025; Yang et al., 2025a), and potentially offering a principled source of dense intermediate reward for reasoning training. Whether reasoning models in fact exploit this signal remains an open empirical question (see §C for limitations and future directions).

AI USE STATEMENT

We used Gemini to assist with coding and with improving the clarity of the writing. All AI-assisted code and text were reviewed by the authors. The authors take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experimental details needed to reproduce our results are provided in the appendices. Models, datasets, sample sizes, and preprocessing are specified in §A.1; full prompts for all three phases, the foil experiment are in §A.2–§A.2.1 and Figure A1. Linear probing follows a fixed protocol (L_2-regularised logistic regression, $C = 0.001$, 5-fold stratified cross-validation, fixed random seed) detailed in §A.4, with P(IK) estimation in §A.3. Causal intervention procedures—patching, mean ablation, calibration-set construction, and the self-correction noising paradigm with its trial-selection rule—are specified in §A.6–§A.8. All generation used greedy decoding (temperature $= 0$) on publicly available models (`google/gemma-3-27b-it`, `Qwen/Qwen2.5-7B-Instruct`) and datasets (TriviaQA, PopQA, MNLI).

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

Leonardo Bertolazzi, Philipp Mondorf, Barbara Plank, and Raffaella Bernardi. The validation gap: A mechanistic analysis of how language models compute arithmetic but fail to validate it. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29375–29412, 2025.

Malcolm W Brown and John P Aggleton. Recognition memory: what are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, 2(1):51–61, 2001.

Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in llms. *Advances in Neural Information Processing Systems*, 37:138393–138431, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Jiefeng Chen, Jie Ren, Xinyun Chen, Chengrun Yang, Ruoxi Sun, Jinsung Yoon, and Sercan Ö Arık. Sets: Leveraging self-verification and self-correction for improved test-time scaling. *arXiv preprint arXiv:2501.19306*, 2025.

Stephen M Fleming and Nathaniel D Daw. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, 124(1):91, 2017.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective STaRs. *arXiv:2503.01307*, 2025.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of language model confidence estimation and calibration. *arXiv preprint arXiv:2311.08298*, 2023.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Tim Tian Hua, Andrew Qin, Samuel Marks, and Neel Nanda. Steering evaluation-aware language models to act like they are deployed. *arXiv preprint arXiv:2510.20487*, 2025.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Adam Kepecs and Zachary F Mainen. A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367 (1594):1322–1337, 2012.

Adam Kepecs, Naoshige Uchida, Hatim A Zariwala, and Zachary F Mainen. Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210):227–231, 2008.

Roozbeh Kiani and Michael N Shadlen. Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, 324(5928):759–764, 2009.

Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Veličković. How do llms compute verbal confidence? In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.17839.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Loka Li, Zhenhao Chen, Guangyi Chen, Yixuan Zhang, Yusheng Su, Eric Xing, and Kun Zhang. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *arXiv preprint arXiv:2402.12563*, 2024.

Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, Ruiyang Qin, Yiyu Shi, et al. Large language models have intrinsic self-correction ability. *arXiv preprint arXiv:2406.15673*, 2024.

Jiarui Liu, Jivitesh Jain, Mona Diab, and Nishant Subramani. Llm microscope: What model internals reveal about answer correctness and context utilization. *arXiv preprint arXiv:2510.04013*, 2025.

Yi-Long Lu, Jiajun Song, and Wei Wang. A unified representation underlying the judgment of large language models. *arXiv preprint arXiv:2510.27328*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*, 2024.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Alexandre Pouget, Jan Drugowitsch, and Adam Kepecs. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, 19(3):366–374, 2016.

Patrick M. A. Rabbitt. Error correction time without external error signals. *Nature*, 212:438, 1966. doi: 10.1038/212438a0.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belém, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7:221–231, 2025.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. *arXiv preprint arXiv:2410.12877*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Base models know how to reason, thinking models learn when. *arXiv preprint arXiv:2510.07364*, 2025.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *ICLR*, 2023.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. Reasoning-finetuning repurposes latent representations in base models. *arXiv:2507.12638*, 2025.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2550–2575, 2023.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1112–1122, 2018.

Xiaohu Xie, Xiaohu Liu, and Benjamin Yao. Know when you're wrong: Aligning confidence with correctness for LLM error detection. *arXiv preprint arXiv:2603.06604*, 2026.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Xiao-Wen Yang, Xiao-Yu Zhu, Wei-Da Wei, De-Chuan Zhang, Jian-Jun Shao, Zhi Zhou, Lan-Zhe Guo, and Yu-Feng Li. Step back to leap forward: Self-backtracking for boosting reasoning of language models. *arXiv:2502.04404*, 2025a.

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. Confidence vs critique: A decomposition of self-correction capability for llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3998–4014, 2025b.

Nick Yeung, Matthew M. Botvinick, and Jonathan D. Cohen. The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, 111(4):931–959, 2004. doi: 10.1037/0033-295X.111.4.931.

Dongryeol Yoon, Seongyun Kim, Sukyung Yang, Seongjin Kim, Yireun Kim, Eunji Kim, Eunsol Choi, Yohan Kim, and Minjoon Seo. Reasoning models better express their confidence. *arXiv preprint arXiv:2505.14489*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

Qingjie Zhang, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, Minlie Huang, Ke Xu, Hewu Li, Liu Yan, and Han Qiu. Understanding the dark side of llms' intrinsic self-correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27066–27101, 2025.

APPENDIX OVERVIEW

- **Appendix A: Methods** (§A)
  - A.1 Models and Datasets (§A.1)
  - A.2 Experimental Paradigm (§A.2)
  - A.3 P(IK): Probability that I Know (§A.3)
  - A.4 Linear Probing (§A.4)
  - A.5 Length-Normalised Answer Log-Probability (§A.5)
  - A.6 Error Detection: Activation Patching (§A.6)
  - A.7 Error Detection: Activation Mean Ablation (§A.7)
  - A.8 Self Correction: Activation Mean Ablation (§A.8)
- **Appendix B: Detailed Review of Literature** (§B) — Detailed review of self-correction, latent representations, mechanistic interpretability, and decision neuroscience literature
- **Appendix C: Limitations of Current Work** (§C) — greedy decoding, scope of paradigm, reasoning models
- **Appendix D: Supplemental Figures** (§D) — Prompts, supplemental probing figures, causal logit difference figures
- **Appendix E: Supplemental Tables** (§E) — Individual predictor AUROCs for verification, answer change, and correctability (Tables A2–A9)
- **Appendix F: Supplemental Results — Gemma 3 27B on TriviaQA**
  - F.1 Verbal confidence predicts error detection above additional log-probability baselines (§F.1)
  - F.2 Error Detection is Graded by Answer Plausibility (§F.2) — Foil experiment behavioural results
  - F.3 P(IK) Analyses (§F.3)
  - F.4 Error Detection Causal Experiment Details (§F.4) — Full patching and ablation results, joint ablation, LAT confound
  - F.5 Self Correction Causal Experiment Details (§F.5)
- **Appendix G: Cross-Model and Cross-Task Generalisation**
  - G.1 MNLI Behavioural Results (§G.1)
  - G.2 MNLI Probing Results (§G.2)
  - G.3 PopQA Behavioural Results (§G.3)
  - G.4 PopQA Probing Results (§G.4)
  - G.5 Cross-Task Transfer (§G.5)
  - G.6 Practical Relevance (§G.6)
  - G.7 Qwen 2.5 7B Behavioural Results (§G.7)
  - G.8 Qwen 2.5 7B Probing Results (§G.8)
  - G.9 Qwen 2.5 7B Causal Experiments (§G.9)

A METHODS

A.1 MODELS AND DATASETS

Models. We study two instruction-tuned language models: Gemma 3 27B (`google/gemma-3-27b-it`; 62 layers, 5376-dimensional residual stream) and Qwen 2.5 7B (`Qwen/Qwen2.5-7B-Instruct`; 28 layers, 3584-dimensional residual stream). Both models were accessed via HuggingFace Transformers and run with greedy decoding (temperature $= 0$) throughout.

Datasets. We use TriviaQA (Joshi et al., 2017) as our primary dataset, a factual question-answering benchmark requiring retrieval of real-world knowledge. After deduplication, our TriviaQA sample comprises 7,227 questions for Gemma and 3,500 for Qwen. We additionally test models on PopQA (Mallen et al., 2023): this comprises 13,068 entity-centric factual questions derived from Wikidata, spanning 16 relation types (e.g. director, screenwriter, capital, country). Questions are templated ("Who was the director of X?") with short entity answers (mean 1.9 words), providing a structurally distinct complement to TriviaQA's diverse, naturally-phrased questions. We also test models on the Multi-Genre Natural Language Inference (MNLI) dataset (Williams et al., 2018), a three-way classification task (entailment, contradiction, neutral), using the development set downloaded from HuggingFace ($n = 9{,}888$) and applied to Gemma only. As entailment and contradiction trials show near-ceiling verification endorsement rates with minimal variance, all MNLI analyses are restricted to neutral trials ($n = 3{,}395$).

A.2 EXPERIMENTAL PARADIGM

Phase 0: Answer generation and confidence rating. In Phase 0, the model was presented with questions and asked to generate a concise answer followed by a verbal confidence rating (see Figure A1 for the full prompt and class definitions). The prompt instructed the model to select one of ten confidence classes spanning the $[0, 1]$ interval (e.g., "No chance" for 0.0–0.1, "Almost certain" for 0.9–1.0). The model was explicitly instructed not to provide reasoning or explanation, yielding short answers typically one to three tokens in length. For MNLI, the same Phase 0 prompt structure was used, with the model generating a classification label (entailment, neutral, or contradiction) and a confidence rating.

Phase 1: Verification. In Phase 1, the model was presented with its own Phase 0 answer and asked to verify its correctness. The model was shown the original question and its prior answer and prompted to judge whether the answer was correct, outputting only a single letter ("Y" or "N") (see Figure 1). The verification prompt contained the model's answer followed by a newline token (NL) and then the verification instruction:

```
Question: {question}
Your answer: {model_answer}
⟨NL⟩                                    ← PANL: activations extracted here
Verify your answer. Correct? (Output ONLY a single letter,
Y/N):
```

The newline token (NL) immediately following the model's answer is the position we term PANL (post-answer newline). Due to causal attention, PANL can attend to all question and answer tokens but not to any subsequent tokens in the verification instruction. It therefore has the potential to reflect a backward-looking evaluative signal relating to question–answer fit. Residual stream activations were extracted at PANL following the methodology established in Kumaran et al. (2026).

For MNLI, the Phase 1 prompt followed the same structure:

```
Question: {question}
Your answer: {model_answer}
Verify your answer. Correct? (Output ONLY a single letter,
Y/N):
```

Phase 2: Self-correction. In Phase 2, the model received the original question, its prior answer, and its own verification response, and was asked to provide what it believed to be the correct answer:

```
Question: {question}
Your answer: {model_answer}
You said: {verification_response}
What do you believe is the correct answer to this question?
```

For MNLI, the Phase 2 prompt was adapted to the classification task:

```
Question: {question}
Your answer: {model_answer}
Verify your answer. Correct?: {verification_response}
What is the correct relationship - entailment, neutral, or
contradiction?
(Output ONLY one of: entailment, neutral, contradiction):
```

We denote the Phase 0 answer as A1 and the Phase 2 answer as A2. The verification response partitions trials into four signal-detection cells: hits (correct answer, verified "Y"), misses (correct answer, verified "N"), correct rejections (incorrect answer, verified "N"), and false alarms (incorrect answer, verified "Y").

Answer scoring. Because A2 responses were typically discursive—often containing reasoning, hedging, or restated context rather than a concise answer—we used GPT-4o to judge the correctness of both A1 and A2 responses against the ground-truth TriviaQA answer. For consistency, GPT-4o was used to score A1 as well, despite A1 answers being short enough for string matching in principle. GPT-4o scoring of A1 (and in fact A2) was highly consistent with exact string matching (agreement rate 93.2% and 83.8% for A1 and A2, respectively). We additionally validated GPT-4o scoring against human judgment on 196 TriviaQA trials, stratified across genuine corrections, failed corrections, A1 edge cases, and a random sample. Overall A2 correctness agreement was 90% ($\kappa = .79$), answer-change agreement was 95% ($\kappa = .89$), and A1 correctness agreement was 88% (98% on the $n = 50$ random subset). We did not separately validate the foil conditions, but the foil A2-correctness results rely on the same GPT-4o scoring pipeline (foils differ from the main experiment only in the answer presented for verification, not in how A2 is scored), so the agreement figures above apply. Together, these checks confirm that the use of a neural judge did not introduce systematic bias.

A.2.1 TriviaQA: Foil Experiment

To test whether the model's error-detection capacity extends beyond self-generated answers, we constructed a foil verification experiment using a subset of 1,929 TriviaQA questions. For each question, we used GPT-4o to generate three types of foil answers: a hard foil (a plausible but incorrect alternative, semantically close to the correct answer and in the same category or domain), an easy foil (a topically related but clearly wrong answer that anyone with basic knowledge should reject), and an unrelated foil (an answer from a completely different domain that is immediately obvious as incorrect). The model was also presented with its own Phase 0 answer as a baseline condition. Each question appeared in all four conditions, enabling within-question comparisons. The verification prompt was identical to the main experiment except that "Your answer" was replaced with "The candidate's answer", removing any indication that the answer was self-generated. Self-correction followed the same procedure as the main experiment. We assessed error-detection sensitivity (d'), response criterion (c), verbal confidence, answer change rate, and correction success for each condition.

A.3 P(IK): Probability that I Know

Following Kadavath et al. (2022), we estimated the model's latent knowledge about each question using sampling consistency. For each question, we generated 20 independent responses at temperature $= 1$ and computed the proportion that matched the ground-truth answer, yielding a question-level estimate of retrieval reliability (P(IK)) with 21 discrete levels ($0/20$ through $20/20$). P(IK) was included in nested regression analyses to ensure that PANL's contribution is not reducible to question-level knowledge.

A.4 Linear Probing

Following Kumaran et al. (2026), we trained L_2-regularised logistic regression probes ($C = 0.001$, 5-fold stratified CV) on z-scored residual stream activations from all 7,227 TriviaQA trials. We extracted activations at four positions: PANL (the first token following the answer), the last answer token (LAT), the prompt's last token, and a control position (third question token). Binary targets were: (i) verification response, (ii) answer change, and (iii) A2 correctness; for the continuous target of verification log-probability difference we used Ridge regression and report Pearson r. To isolate the contribution of PANL beyond behavioural signals, we compare against a baseline logistic regression trained on A1 correctness, mean answer log-probability, verbal confidence, and verification log-probability difference. To obtain a scalar summary of PANL information for use in downstream analyses, we extracted each trial's cross-validated predicted probability (i.e., the prediction from the fold in which it was held out); we refer to this as the *probe score* and enter it as a predictor alongside behavioural measures in logistic regressions and AUROC comparisons.

Surface feature controls for linear probing. To rule out the possibility that linear probes exploit shallow textual statistics rather than genuine internal representations, we constructed a surface-feature baseline comprising TF-IDF lexical features (100 features each for question and answer text, fit separately) and token lengths of both question and answer. A logistic regression trained on these 202 surface features achieved near-chance prediction of verification response (AUROC $= .564$), answer change ($.564$), and A2 correctness ($.585$), and added no predictive value when appended to the behavioural scalar baseline ($p = 1.0$ for all targets). This confirms that the probing results reported above cannot be attributed to surface properties of the input or output text.

A.5 Length-Normalised Answer Log-Probability

To obtain a white-box measure of model confidence, we extracted token-level log-probabilities during answer generation. For each generated sequence, we computed the log-probability of each token t_i given the preceding context:

$$\log p(t_i \mid t_{<i}, \mathbf{x}) \tag{1}$$

where $\mathbf{x}$ denotes the input prompt. We then computed the mean log-probability over the n answer tokens:

$$\bar{\ell} = \frac{1}{n} \sum_{i=1}^{n} \log p(t_i \mid t_{<i}, \mathbf{x}) \tag{2}$$

This length-normalised metric controls for variation in answer length, ensuring that longer answers are not penalised simply for having more tokens. Answer token boundaries were identified by mapping the extracted answer string back to the generated token sequence.

A.6 Error Detection:Activation Patching

Corruption of answer tokens via mean ablation. To test whether specific position–layer combinations are causally sufficient for error detection, we use a corrupt-and-restore procedure following Meng et al. (2022); Heimersheim & Nanda (2024); Zhang & Nanda (2023); Wang et al. (2023). We first disrupted the model's access to answer information through mean ablation of answer tokens. Let $\mathbf{x}_i^{(0)}$ denote the embedding of token i in the input sequence, and let $\mathcal{A} = \{a_1, \ldots, a_k\}$ denote the set of answer token positions in the verification prompt. We computed mean activations from a calibration set $\mathcal{C}$ of 100 trials (disjoint from the test set):

$$\bar{\mathbf{x}}_j^{(0)} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbf{x}_{j,c}^{(0)} \tag{3}$$

where $\mathbf{x}_{j,c}^{(0)}$ is the embedding at answer position j for calibration trial c. For each test trial, we replaced all answer token embeddings with their corresponding mean activations, propagating this corruption through the entire forward pass. This mean ablation effectively destroys the correspondence between the question and the specific answer presented for verification, replacing answer-specific information with a position-averaged signal that retains no trial-level content.

Patching procedure. Let $\mathbf{h}_p^{(\ell)}$ denote the residual stream activation at position p after layer ℓ. For a given test trial, let $\mathbf{h}_p^{(\ell,\text{clean})}$ denote the activation from a clean forward pass (no corruption) and $\mathbf{h}_p^{(\ell,\text{corrupt})}$ denote the activation when answer tokens have been mean-ablated. Our patching intervention selectively restores the clean activation at position p and layer ℓ:

$$\mathbf{h}_p^{(\ell,\text{patched})} = \mathbf{h}_p^{(\ell,\text{clean})} \tag{4}$$

while all other positions retain their corrupted activations. This intervention was applied after the MLP block at each layer. We tested five positions: the last answer token (LAT), PANL (post-answer newline), PANL+1, PANL offset 9, and the last token of the verification prompt. We evaluated rescue of error detection using two measures: d' (sensitivity to errors) and the verification logprob difference ($\log P(\text{Y}) - \log P(\text{N})$).

A.7 Error Detection: Activation Mean Ablation

To test whether representations at specific token positions are *necessary* for error detection, we performed mean ablation experiments. For each position of interest, we replaced its residual stream activation at a target layer with the mean activation computed from a balanced calibration set of 200 trials (50 per SDT cell: hit, miss, false alarm, correct rejection). This calibration set was disjoint from the test set. Balancing across SDT cells ensures that the mean activation is not biased toward any particular verification outcome, so the replacement removes trial-specific information without injecting a systematic signal.

We note that mean ablation does not push the model toward a semantically meaningful "neutral" verification state. The mean of activations encoding correct and incorrect evaluations is not itself an encoding of intermediate confidence—analogous to how averaging word embeddings for "brilliant" and "terrible" does not yield a representation of "mediocre." Rather, mean ablation disrupts the position's contribution to the computation by replacing trial-specific information with an uninformative average.

We tested five positions ($n = 1{,}000$ trials per layer per position): the last answer token (LAT), PANL (post-answer newline), PANL+1, PANL offset 9, and the last token of the verification prompt. For each position, we applied mean ablation at individual layers across the network and measured the resulting disruption to verification behaviour. We assessed disruption using two measures: d' (error-detection sensitivity) and the verification logprob difference ($\log P(\text{Y}) - \log P(\text{N})$). If a position is necessary for error detection, ablating it should reduce d' toward chance and shift the logprob difference. If ablation has no effect, the position is not necessary for the model's error-detection mechanism.

A.8 SELF-CORRECTION: ACTIVATION MEAN ABLATION

We adapted the mean-ablation paradigm used in the verification experiment to test whether the same residual-stream positions drive self-correction. The prompt presents the question, A_1 (the model's initial answer), a forced verification response ("Is your answer correct (Y/N)? N"), and a request for a final answer (see exact prompt below); the assistant turn is prefilled with "Final answer:" so that the next-token position evaluates over candidate A_2 tokens. The dependent measure is $P(\text{GT})$ at this position—the proportion of trials where the model's argmax output is the first token of the ground-truth answer.

The test set was first restricted to trials with demonstrated correction capacity: A_1 incorrect and A_2 correct under the original two-step Phase 2 pipeline. Trials were then selected based on if the clean baseline of the simplified prompt (shown below) elicited correction ($\log P(\text{GT}) > \log P(A_1)$): $n = 122$).

Correction prompt for Causal Experiment. The simplified one-step correction prompt is structured as follows, with {`question`} and {`A1`} replaced by trial-specific values and "Final answer:" prefilled in the assistant turn so that the next generated token begins the corrected answer:

```
Question: {question}
Your answer: {A1}

Is your answer correct (Y/N)? N
What is your final answer? Output ONLY the answer, with
no explanation.
```

PANL is the newline token immediately following {`A1`}, and LAT is the final token of {`A1`}.

Mean activations for ablation were computed over a balanced 200-trial calibration set (50 per SDT cell). Interventions were applied at four positions—last answer token (LAT), post-answer newline (PANL), PANL+1 (position-specific control), and joint LAT+PANL.

B Detailed Review of the Literature

Self-Correction in LLMs. Huang et al. (2024) define intrinsic self-correction as when an LLM tries to correct errors using its inherent abilities. Self-correction more generally has wider definitions including provision of external feedback (extrinsic self-correction) — reviewed by Kamoi et al. (2024). Huang et al. show using a particular prompt that LLMs can't correct reasoning errors and can even cause performance decreases. Zhang et al. (2025) ask models "are you sure, think again" and show that performance decreases. Self-Refine (Madaan et al., 2023) iteratively refines the output of the model until verified as correct, showing improvement, as do Liu et al. (2024); Chen et al. (2025). Self-Enhanced Test-Time Scaling (SETS; Chen et al. 2025) combines parallel sampling with iterative self-verification and self-correction using a single LLM to show increased performance in reasoning tasks. Weng et al. (2023) show that self-verification of reasoning enhances performance. Self-verification is a diagnostic step where the model judges whether a proposed solution satisfies the task constraints, producing a verdict. Self-correction is the subsequent generative step where, conditioned on that verdict and the history of prior attempts, the model produces an improved solution.

Li et al. (2024) look at *intrinsic* self-correction. They use an IoE prompt — if you are very confident in your answer maintain it, otherwise update. They show small gains over a standard prompt. Yang et al. (2025b) show that fine-tuning can improve self-correction, but this could be because the model learns a high-confidence–stay and low-confidence–switch policy. Xie et al. (2026) extract normalized confidence scores from self-evaluation token probabilities (Yes/No) and show that SFT preserves calibration while RL and DPO induce overconfidence through reward exploitation. However, their analysis does not examine whether internal representations encode confidence signals prior to self-evaluation, nor whether such signals predict downstream self-correction beyond behavioral output measures.

Bertolazzi et al. (2025) use circuit analysis to study arithmetic error detection in small LLMs (1.5B–3B), finding that models rely on "consistency heads" — attention heads performing surface-level digit matching — and that validation circuits fire before arithmetic computation completes in higher layers. Their work studies artificially introduced errors on a constrained task and identifies the responsible circuit components. Our approach differs in studying naturally occurring errors on open-domain factual QA, using probes to characterise what *information* the residual stream encodes rather than which *components* comprise the circuit.

Latent Representations of Uncertainty and Correctness. A growing body of work demonstrates that LLMs encode information about the quality of their outputs within internal activations, often in ways that diverge from surface-level confidence. Azaria & Mitchell (2023) showed that models encode a latent representation of truthfulness in their hidden states: classifiers trained on these activations distinguish true from false statements with 60–80% accuracy, generalizing across topics and outperforming prompting-based methods. Critically, classifiers trained on these internal activations outperform methods based on output probabilities, which are confounded by factors such as sentence length and token frequency and thus provide less reliable signals of truthfulness. Similarly, Liu et al. (2025) found that activations at the first output token predict answer correctness with approximately 75% accuracy, and introduced metrics to distinguish correct, incorrect, and irrelevant retrieval contexts directly from model internals. These findings align with broader evidence that task-relevant information may be decodable from network activations yet remain dissociated from observed behavior (Burns et al., 2022; Li et al., 2023; Bürger et al., 2024; Lu et al., 2025).

Orgad et al. (2024) probe LLM hidden states to decode output correctness, finding that the truthfulness signal is concentrated at exact answer tokens — consistent with our finding that token position matters critically for error-related decoding. They taxonomise errors by sampling consistency (e.g., consistently wrong vs. occasionally wrong) and show these types are predictable from internal representations, revealing that models can encode the correct answer while consistently generating an incorrect one. Our work differs in probing a graded confidence signal within a verification paradigm and linking it to error detection and self-correction, rather than decoding binary correctness.

Confidence Calibration. A large literature studies the calibration of LLM confidence: how well an expressed confidence estimate tracks the probability that the first-pass answer is correct, measured via ECE and AUROC against correctness, and how calibration depends on elicitation—token log-probabilities (Kadavath et al., 2022), sampling-based consistency (Tian et al., 2023; Geng et al.,

2023), verbal reports (Xiong et al., 2023; Steyvers et al., 2025; Yoon et al., 2025), and self-evaluation probabilities (Xie et al., 2026). Our work relates to this literature in three ways. First, it extends the target of calibration. Calibration work almost universally treats first-pass correctness as the criterion; we show that confidence signals differ sharply in what they predict about the model's *subsequent self-monitoring*—whether it will flag its answer as wrong, whether it will revise, and whether revision will succeed. These targets dissociate from first-pass calibration: verbal confidence is reasonably calibrated for A1 correctness (ECE $= 0.13$) yet carries no information about correction success (AUROC $= .524$), so a well-calibrated first-pass score can be uninformative about precisely the outcomes self-correction depends on. Second, the second-order framework gives the different confidence measures distinct theoretical status rather than treating them as interchangeable estimators: log-probabilities index the generation signal (X_{act}), verbal confidence is a partial, lossy readout of the evaluative signal (X_{eval}), and PANL activations carry that signal directly. This offers a computational explanation for why elicitation methods disagree (Xiong et al., 2023) and why elicitation strategy affects calibration (Tian et al., 2023): different methods tap different underlying quantities. Third, because verbal confidence is a lossy readout of a richer internal signal, internal-state estimates upper-bound elicited confidence—connecting calibration to the latent-knowledge literature above, and locating where and when the calibratable signal resides: cached at PANL, before any confidence query or verification instruction is presented.

Mechanistic Interpretability: Activation Steering. Activation steering is a technique for causally intervening on model behavior by adding or subtracting directions in activation space. Turner et al. (2023) demonstrated that abstract concepts such as love or hate are encoded as approximately linear directions in transformer representations. These directions can be recovered through various methods, including contrasting mean activations between conditions that differ along a conceptual dimension. Subsequent work has applied steering to modify instruction-following behavior (Stolfo et al., 2024), reasoning (Venhoff et al., 2025), sycophancy and other traits (Panickssery et al., 2023), and evaluation-aware responses (Hua et al., 2025).

Activation Patching. Activation Patching (also termed causal tracing) identifies which model components are causally responsible for specific behaviors. Meng et al. (2022) used the corrupt-then-restore paradigm: inputs are corrupted to disrupt model output, then clean activations are selectively restored at specific position-layer combinations to measure recovery. Positions that restore performance when patched are implicated as causally sufficient for the computation. This approach has been extended to study indirect object identification (Wang et al., 2023), and best practices for activation patching metrics have been systematically evaluated (Zhang & Nanda, 2023; Heimersheim & Nanda, 2024).

Theories of Confidence in Decision Neuroscience. Our investigation connects to a longstanding debate in decision neuroscience concerning the computational basis of confidence (Fleming & Daw, 2017; Pouget et al., 2016; Kepecs & Mainen, 2012; Kiani & Shadlen, 2009). Under first-order accounts, confidence arises from the same internal signals that drive the decision itself. In perceptual tasks, for instance, both the choice and confidence derive from a single decision variable representing accumulated evidence; confidence is simply a readout of how strongly this variable favored the chosen option. Translated to LLMs, a first-order account would hold that verbal confidence is a readout of token log-probabilities—the same signals that determined which answer tokens to generate also determine confidence.

Under second-order accounts, confidence involves signals that are distinct from—though correlated with—those driving the decision (Fleming & Daw, 2017). These additional signals enable an evaluation of the decision that goes beyond the information directly used to produce it. For LLMs, evidence that verbal confidence reflects information beyond token log-probabilities would suggest a second-order-like computation capable of genuine answer-quality evaluation.

An important consequence is that second-order architectures can support error detection: because confidence draws on partially independent information, the system can recognize that a response may be wrong even after committing to it. In contrast, pure first-order architectures cannot detect errors, because confidence and decision accuracy are yoked to the same underlying signal.

Recall, Recognition, and the Computational Basis of Error Detection The distinction between generation-time signals and the evaluative signal encoded at PANL has a natural analogue in the memory literature. In episodic memory, recall and recognition dissociate both behaviourally and neurally: recall requires active reconstruction of a memory trace from a retrieval cue, whereas recognition can proceed via familiarity signals that do not depend on successful recollection (Brown & Aggleton, 2001). An agent can fail to recall something it can nonetheless recognise as wrong or unfamiliar, because the two processes draw on partially independent computational substrates.

The mechanistic basis for why generation-time signals in LLMs are analogous to recall is provided by work on factual retrieval. Meng et al. (2022) demonstrated that factual associations are localised in MLP weight matrices at specific layers, with answer generation corresponding to a key-value lookup in which the subject token representation activates a stored association to produce the predicted output. On this view, token log-probabilities index the strength of parametric retrieval—how cleanly the subject-to-fact association fired—rather than its accuracy. A confidently retrieved but incorrect association yields high log-probability and, in a purely first-order system, provides no signal that an error has occurred.

The PANL representation is structurally positioned to perform a different computation. As established in Kumaran et al. (2026), PANL attends backward over the completed answer and encodes a summary of question–answer fit that explains substantial variance in verbal confidence beyond token log-probabilities. This backward-looking integration over a finished object—rather than a forward lookup through parametric memory—is precisely what distinguishes recognition from recall: the evaluative process is not hostage to whether retrieval succeeded, because it operates on the output of retrieval rather than performing it.

Taken together, the Meng et al. mechanistic account and the Brown & Aggleton computational framework provide complementary explanations for why generation-time and evaluative signals dissociate. The former specifies what generation consists of in LLMs; the latter provides the principled computational reason why a post-hoc evaluation over the completed response should have access to information that the generation signal does not.

C Limitations of Current Work

We use greedy decoding throughout. This is by design: it guarantees that A_1 is the argmax of the model's own token distribution, so any successful revision cannot come from stochasticity at resampling—it must come from the evaluative process reweighting the alternatives. But it means that it remains for future work to determine how the PANL signal behaves over sampled answers, where the committed response is not the mode. The foil experiment offers an insight: the model detects errors in answers it did not generate (i.e. they were produced by prompting of GPT4o to produce foils of varying difficulty) , and detection scales with how wrong the answer is. Foils, therefore, came from a different model, not from the model's own sampling distribution. Future work should test whether PANL carries the same signal when answers are sampled at temperature.

We study base instruction-tuned models, not reasoning models. Our motivation was that factual QA without chain-of-thought is the acutest test of first-order versus second-order accounts. With a visible reasoning trace, a model can find its mistake by re-reading its own steps—error detection can work over the trace, without any signal beyond generation. With no trace, that route is closed. The only way to detect an error is to evaluate question–answer fit directly—recognition rather than recall (see §B)—and this evaluation is precisely what separates a second-order system from a first-order one. Whether reasoning models use the PANL signal is a natural next question.

Our paradigm has limits of scope. The tasks are short-answer factual QA and three-way NLI. Whether the same evaluative signal operates in other domains—code, mathematics, multi-step reasoning, open-ended generation—is untested. These domains differ in ways that matter: errors in code can be checked against execution, errors in reasoning against intermediate steps, and long-form answers have no single token position where an evaluative summary would naturally consolidate. Whether PANL-like signals exist in these settings, and where they live, is a question for future work.

D Supplemental Figures

```
"""First, provide an answer to the question: do NOT give any REASONING or EXPLANATION.
Then, classify your confidence into one of the following classes based on how likely
your answer is to be correct (again NO REASONING OR EXPLANATION):

- "No chance" (0.0–0.1)
- "Really unlikely" (0.1–0.2)
- "Chances are slight" (0.2–0.3)
- "Unlikely" (0.3–0.4)
- "Less than even" (0.4–0.5)
- "Better than even" (0.5–0.6)
- "Likely" (0.6–0.7)
- "Very good chance" (0.7–0.8)
- "Highly likely" (0.8–0.9)
- "Almost certain" (0.9–1.0)

Each category reflects the probability that the answer is correct.

At the very end of your output, format your confidence as
**Confidence**: $CLASS
where CLASS is one of the names (only the names without the probability ranges)
of the classes above.

Question: {question}:"""
```

Figure A1: Full categorical confidence class prompt used in Phase 0. This prompt is derived from (Yoon et al., 2025) but with the following key modification: the first token of every confidence class is unique, allowing us to meaningfully analyze changes in the ID of first token, the logit of the first token etc.

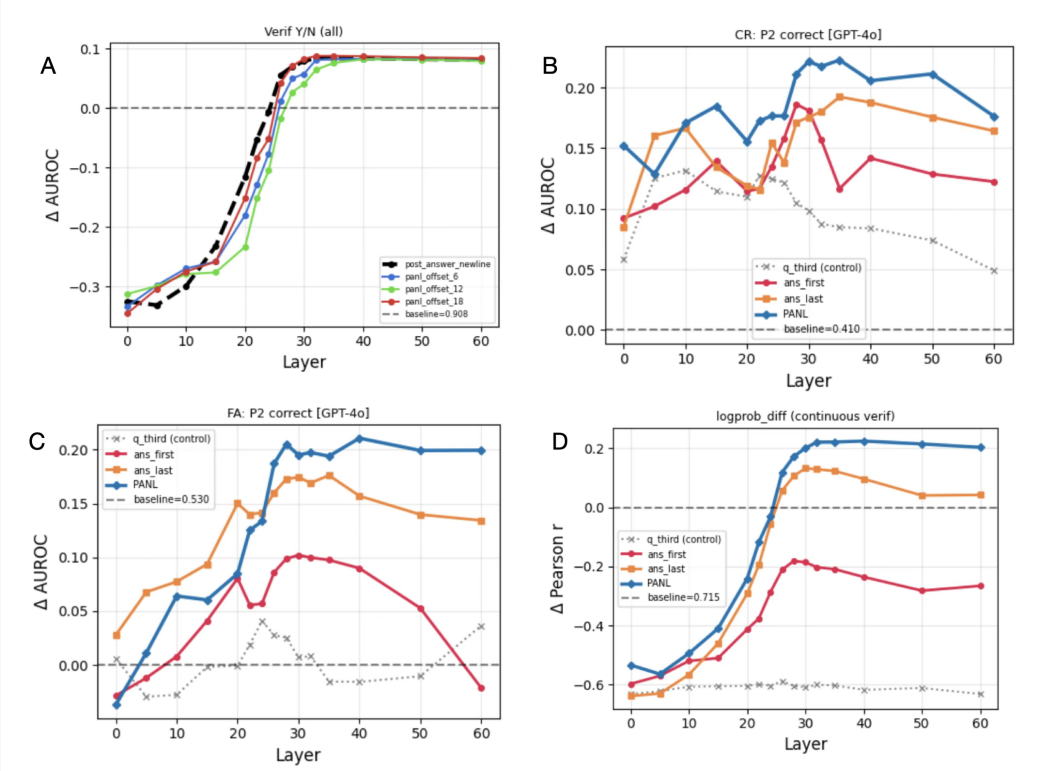

Figure A2: Supplementary linear probe results. Increase in AUROC (*delta above behavioral baseline*) shown for PANL and several downstream positions. **A**: Verification prediction from PANL, first answer token, last answer token, and downstream positions (offset 6, 12, 18 tokens after PANL). **B**: Predicting A2 correctness within correct rejection (CR) trials. **C**: Predicting A2 correctness within false alarm (FA) trials. **D**: Prediction of verification confidence (logprob_diff) as a continuous variable (Δ Pearson r above behavioral baseline).

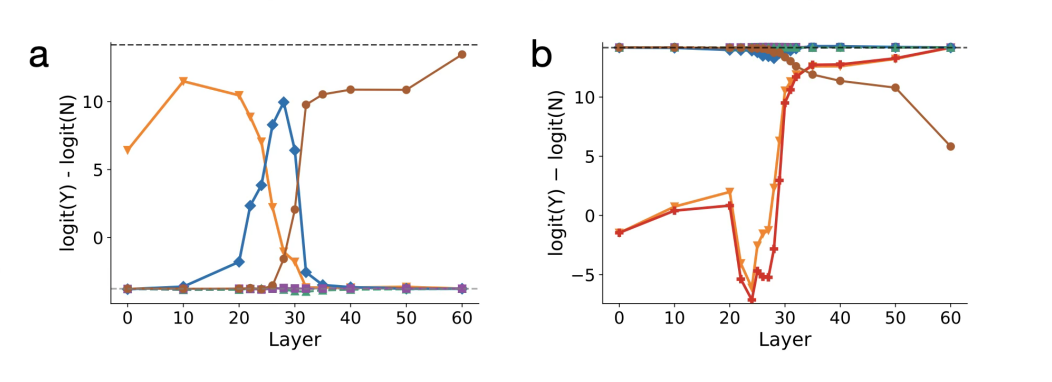

Figure A3: Gemma 3 27B causal experiments: verification logprob difference ($\log P(\mathrm{Y}) - \log P(\mathrm{N})$). **a**: Activation patching. Clean baseline: 14.20 (black dashed); corrupted: -4.15 (grey dashed). The same three positions that rescue d' (Figure 5a) also rescue the graded verification signal, with the same temporal dissociation: LAT (orange) at early layers, PANL (blue) at mid layers, prompt last token (brown) at later layers. PANL+1 and PANL offset 9 show no rescue. **b**: Activation mean ablation. LAT ablation (orange) produces large negative shifts at early-to-mid layers (peak $\Delta = -20.3$ at L24); joint LAT+PANL ablation (red) produces additional shifts at mid layers, consistent with redundancy. PANL alone (blue) shows no reliable shift. Prompt last token ablation (brown) disrupts at later layers.

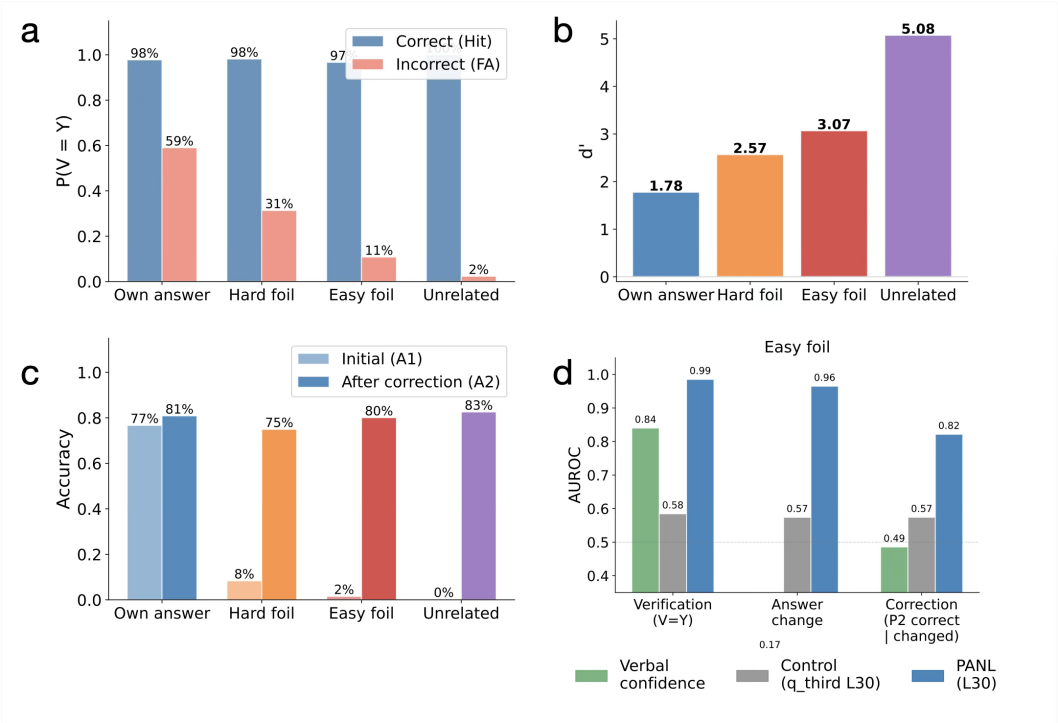

Figure A4: Foil experiment results ($n = 1{,}929$ questions per condition, same questions across all). **a**: Verification endorsement rate (V=Y) for correct answers (hit rate, blue) and incorrect answers (false alarm rate, red) by condition. Hit rates remain near ceiling ($\geq 97\%$) across all conditions, while false alarm rates decrease monotonically from own answers (59%) through hard foils (31%), easy foils (11%), to unrelated foils (2%). **b**: Error-detection sensitivity (d') increases with foil discriminability, from 1.78 (own answer) to 5.08 (unrelated foil). The response criterion shifts from strong Y-bias ($c = -1.14$, own answer) toward neutrality ($c = -0.14$, unrelated), indicating that the self-affirmation bias observed in the main experiment is specific to self-generated answers. **c**: Accuracy before (A1, light) and after (A2, dark) self-correction. For foil conditions, A1 accuracy reflects the foil's incorrectness (0–8%); after correction, accuracy rises to 75–83%, demonstrating that the model can recover the correct answer for the majority of questions. **d**: PANL L30 probe transfer to foil conditions. Probes trained on own-answer trials predict verification (AUROC = .84), answer change (.99), and correction success (.82) on easy foil trials, with the control position (grey) at or below chance. As in the main experiment, verbal confidence barely predicts answer change success above chance. See Table A5 for results for other foils (AUROC).

E SUPPLEMENTAL TABLES

Table A1: PANL probe score improves prediction of verification, answer change, and correction success beyond the best behavioural baseline. *Behav* is an L_2-regularised logistic regression on all available behavioural predictors (see Tables A2–A4 for individual predictor AUROCs). *PANL* is the cross-validated probe score from layer 30 activations at the post-answer newline position. The likelihood-ratio test assesses whether PANL adds to the combined behavioural model. Answer change and A2 correctness are restricted to incorrect trials; A2 correctness is further restricted to trials where the model changed its answer, the subset where the prediction task is non-trivial (correct trials are almost exclusively hits that do not change).

Target	n	**Behav**	**PANL**	**Behav + PANL**	**LR χ^2**	p
Verification Y/N (all trials)	7,223	.908	.986	.983	1100.0	$< .001$
Answer change (incorrect)	1,764	.901	.921	.931	237.7	$< .001$
A2 correct (incorrect + changed)	856	.475	.614	.605	9.5	$< .01$

Table A2: Predicting verification response (V=Y vs V=N): individual and combined predictor AUROCs. *Combined* is an L_2-regularised logistic regression on all behavioural predictors listed. PANL is the cross-validated probe score at layer 30. The likelihood-ratio test assesses whether PANL adds to the combined behavioural model. A1 correctness is excluded from the incorrect-only analysis (constant within subset).

		Individual predictors							
Subset	n	**Logprob**	**Conf**	**A1 corr**	**Combined**	**PANL**	**Comb + PANL**	**LR χ^2**	p
All trials	7,223	.665	.800	.843	.908	.986	.983	1100.0	$< .001$
Incorrect	1,764	.530	.703	—	.715	.958	.948	443.7	$< .001$

Table A3: Predicting answer change within incorrect trials. Verification log-probability difference ($\log P(\mathrm{Y}) - \log P(\mathrm{N})$) is the strongest single behavioural predictor; PANL adds significant variance beyond the combined behavioural baseline.

		Individual predictors							
Subset	n	**Logprob**	**Conf**	**Verif LD**	**Combined**	**PANL**	**Comb + PANL**	**LR χ^2**	p
Incorrect	1,764	.508	.657	.907	.901	.921	.931	237.7	$< .001$

Table A4: Predicting A2 correctness within incorrect trials. Within all incorrect trials, verification log-probability difference is a reasonable predictor (AUROC $= .740$), but within the subset that actually changed their answer—where the question is non-trivial—behavioural predictors drop to chance. PANL retains predictive value in this critical subset ($p < .01$). Note that the combined model (Comb + PANL) is slightly below PANL alone in the changed subset: with behavioural features at chance, the combined model is effectively the PANL score plus noise from uninformative covariates, which slightly degrades performance on this smaller sample ($n = 856$).

		Individual predictors							
Subset	n	**Logprob**	**Conf**	**Verif LD**	**Combined**	**PANL**	**Comb + PANL**	**LR χ^2**	p
Incorrect	1,764	.526	.567	.740	.724	.751	.760	46.4	$< .001$
Incorrect + changed	856	.518	.524	.531	.475	.614	.605	9.5	$< .01$

Table A5: Predicting verification response, answer change, and correction success in the foil experiment. *Conf*: verbal confidence alone. *Verif LD*: verification log-probability difference ($\log P(\text{Y}) - \log P(\text{N})$) alone. *Behav*: L_2-regularised logistic regression on all available behavioural predictors (confidence for verification; confidence and verification LD for answer change and correction). *PANL*: cross-validated probe score at layer 30. Correction = A2 correct among trials where the model changed the candidate's answer. A1 correctness is excluded from all analyses as it is constant within each foil condition (always incorrect). The likelihood-ratio test assesses whether PANL adds to the combined behavioural model.

			Individual						
Target	**Condition**	n	**Conf**	**Verif LD**	**Behav**	**PANL**	**Behav + PANL**	**LR χ^2**	p
Verification Y/N	Hard foil	1,924	.886	—	.879	.981	.977	442.1	$< .001$
	Easy foil	1,924	.840	—	.827	.985	.983	258.0	$< .001$
	Unrelated	1,929	.775	—	.771	.995	.994	84.8	$< .001$
Answer change	Hard foil	1,924	.842	.930	.922	.938	.941	165.8	$< .001$
	Easy foil	1,924	.832	.973	.967	.965	.972	74.4	$< .001$
	Unrelated	1,929	.849	.994	.993	.989	.995	18.8	$< .001$
A2 correct | changed	Hard foil	1,444	.510	.656	.637	.859	.852	67.5	$< .001$
	Easy foil	1,791	.514	.697	.694	.822	.820	118.0	$< .001$
	Unrelated	1,915	.519	.711	.706	.813	.805	122.6	$< .001$

Table A6: Cross-model and dataset table for predicting verification response and A2 correctness AUROC: behavioural baseline vs. PANL linear probe. Behavioural baseline = L2-regularised logistic regression on available behavioural predictors (answer logprob, verbal confidence, A1 correctness for verification; answer logprob, verbal confidence, verification logprob difference for A2 correctness). PANL = cross-validated probe at peak layer. MNLI behavioural baseline uses verbal confidence alone.

	Gemma 3 27B						Qwen 2.5 7B	
	TriviaQA		PopQA		MNLI		TriviaQA	
	Behav	PANL	Behav	PANL	Behav	PANL	Behav	PANL
Verif (all)	.908	**.986**	.802	**.993**	.534	**.868**	.860	**.961**
Verif (incorrect)	.715	**.958**	.732	**.991**	.567	**.813**	.646	**.917**
A2 corr (inc+chg)	.475	**.614**	.671	**.785**	.512	**.862**	.575	**.679**

Table A7: Qwen 2.5 7B: PANL probe prediction of verification and correction success. Format follows Table A1. *Behav* is an L_2-regularised logistic regression on available behavioural predictors: answer log-probability, verbal confidence, and A1 correctness for verification (all); answer log-probability and verbal confidence for verification (errors only); answer log-probability, verbal confidence, and verification logit difference for correction success. *PANL* is the cross-validated probe AUROC from layer-20 activations at the post-answer newline position.

Target	n	**Conf**	**Behav**	**PANL**
Verification (all)	3,500	.703	.860	.961
Verification (errors only)	1,452	.655	.646	.917
Correction success (incorrect)	1,245	.461	.568	.774
Correction success (inc. + changed)	741	.528	.575	.679

Table A8: Cross-dataset transfer of PANL L30 probes (Gemma 3 27B). TriviaQA ↔ PopQA share the factual QA domain but differ in question format (diverse vs. templated); TriviaQA ↔ MNLI differ in both domain and answer format (factual QA vs. NLI). Within-dataset values are 5-fold cross-validated AUROC; transfer values apply a probe trained on all source-dataset trials to all trials of the target dataset, so transfer AUROC can exceed the cross-validated within-task value. Verification probes transfer near-perfectly between factual QA datasets but weakly from TriviaQA to MNLI; correctability transfer follows the same ordering at lower levels, with low cosine similarity between correctability directions throughout.

	TriviaQA ↔ PopQA				TriviaQA ↔ MNLI			
Target	Within T	Transfer T→P	Within P	Transfer P→T	Within T	Transfer T→M	Within M	Transfer M→T
Verif (all)	.986	.982	.993	.988	.986	.587	.868	.934
Verif (incorrect)	.958	.977	.991	.961	.958	.573	.813	.897
P2 corr (inc+chg)	.614	.663	.785	.610	.614	.547	.862	.477
Cosine similarity between probe weight vectors								
Verif (all)		.397				.132		
P2 corr (inc+chg)		.071				.027		

Selection rule	% routed	Precision	Recall	Accuracy
TriviaQA (n = 7,207, A_1 accuracy = 75.7%)				
Keep A_1 (no correction)	0%	–	0.0%	75.7%
Use A_2 (unconditional correction)	100%	4.1%	100.0%	79.3%
Correct when V=N	8.6%	28.7%	60.2%	78.0%
PANL probe, top 10%	10%	25.7%	62.9%	78.1%
PANL probe, top 50%	50%	8.0%	97.6%	**79.2%**
PopQA (n = 12,986, A_1 accuracy = 29.4%)				
Keep A_1 (no correction)	0%	–	0.0%	29.4%
Use A_2 (unconditional correction)	100%	3.0%	100.0%	31.7%
Correct when V=N	18.1%	9.9%	60.1%	30.8%
PANL probe, top 10%	10%	16.2%	54.4%	30.7%
PANL probe, top 50%	50%	5.7%	94.8%	**31.6%**

Table A9: **Selective self-correction via PANL probe (Gemma 3 27B).** Setup: a 5-fold cross-validated logistic regression probe is trained on PANL activations at layer 30 to predict "switch helps" (= A_1 wrong AND A_2 correct); ground truth and A_2 outcomes are used only to construct training labels, not at deployment. At deployment, the probe scores each trial from its PANL activation alone, at the end of the first-pass answer A_1. Trials scoring above the threshold are routed to the self-correction prompt and take the corrected answer A_2 as their final answer; the remaining trials retain A_1. Final accuracy is computed across all trials. Baselines: *Keep* A_1 never invokes correction; *Use* A_2 (unconditional self-correction, no probe) invokes correction on every trial; *Correct when V=N* routes trials based on the model's own verification token. The PANL probe is shown at two thresholds (top 10% and top 50% of trials by probe score). Columns: *% routed* = fraction of trials sent to the correction prompt (a proxy for additional inference cost); *Precision* = fraction of routed trials where switching from A_1 to A_2 actually helps; *Recall* = fraction of all helpful switches captured; *Accuracy* = overall correctness across all n trials. Base rate of "switch helps" is 4.1% (TriviaQA) and 3.0% (PopQA); probe AUROC for predicting "switch helps" is 0.90 and 0.86 respectively. The PANL probe at top 50% recovers the unconditional-correction accuracy within 0.1 percentage points while invoking the self-correction prompt on only half of all trials—a 2× reduction in correction-time inference cost with negligible accuracy loss.

F SUPPLEMENTAL RESULTS: GEMMA 3 27B ON TRIVIAQA

F.1 VERBAL CONFIDENCE PREDICTS VERIFICATION DECISION ABOVE ADDITIONAL LOG-PROBABILITY BASELINES

To ensure that the dissociation between verbal confidence and log-probabilities does not depend on summarising the generation signal by the mean alone, we repeated the verification analyses with an enriched set of generation-time statistics: mean, minimum, and first-token answer log-probability, and answer length in tokens. Across all trials, the enriched set improves on mean log-probability only marginally (AUROC .652 $\rightarrow$ .657; LR $\chi^2 = 59.4$), and verbal confidence adds substantially above the full set (LR $\chi^2 = 789.0$, $p < 10^{-173}$). Within incorrect trials, the enriched set adds nothing over the mean (AUROC .521 $\rightarrow$ .537; LR $\chi^2 = 3.1$, $p = .37$), while confidence again adds strongly ($\chi^2 = 161.8$, $p < 10^{-36}$). The second-order signature is thus robust to richer first-order baselines: no combination of chosen-token statistics captures the evaluative information carried by verbal confidence.

F.2 ERROR DETECTION IS GRADED BY ANSWER PLAUSIBILITY

To test whether the model's error-detection capacity extends beyond self-generated answers, we presented the model with foil answers—wrong answers it did not produce (termed "the candidate's answer"; see §A.2.1)—and asked it to verify them using the same paradigm. We compared three foil conditions, each using the same 1,929 questions: (i) a hard foil (a plausible but incorrect alternative, generated by GPT-4o), (ii) an easy foil (a topically related but clearly wrong answer), and (iii) an unrelated foil (an answer to a different question entirely).

Error-detection sensitivity increases monotonically with foil discriminability (Figure A4a,b): $d' = 2.57$ for hard foils, 3.07 for easy foils, and 5.08 for unrelated foils (cf. $d' = 1.78$ for the model's own incorrect answers). The model's evaluation is thus graded by the plausibility of the presented answer, not merely by whether the answer was self-generated.

Notably, the strong Y-bias observed in the main experiment is not a fixed property of the model's verification behaviour. The response criterion shifts toward neutrality as foil plausibility decreases: $c = -0.78$ for hard foils, -0.33 for easy foils, and -0.14 for unrelated foils (cf. $c = -1.14$ for own answers). False alarm rates drop accordingly—from 31% for hard foils to 11% for easy foils and 2% for unrelated foils—while hit rates remain near ceiling ($\geq 97\%$) across all conditions. The Y-bias thus reflects the difficulty of the evaluation rather than a fixed tendency toward endorsement: when the presented answer is clearly wrong, the model readily rejects it.

Verbal confidence tracks this pattern: mean confidence is 0.31 for hard foils, 0.12 for easy foils, and 0.06 for unrelated foils, indicating that the model appropriately calibrates its confidence to the plausibility of the answer under evaluation. Self-correction rates mirror the verification signal: the model changes its answer on 75% of hard foil, 93% of easy foil, and 99% of unrelated foil trials. Importantly, as in the main experiment, verbal confidence barely predicts answer change success above chance (see Figure A4 and Table A5 for results for all foil types). Conditional on changing, correction success is high across all foil conditions (83–89%), consistent with the model having generated the correct answer on its first attempt for the majority of these questions (Figure A4c).

F.3 P(IK) ANALYSES

Within the subset of trials with P(IK) available ($n = 1{,}800$; 411 incorrect), the PANL probe score adds above a baseline of A1 correctness, mean answer log-probability, verbal confidence, and P(IK) for predicting verification on all trials (AUROC = .978 for PANL alone vs .909 for the maximal baseline; LR $\chi^2 = 154.3$, $p < 10^{-34}$) and within incorrect trials (.936 vs .759; LR $\chi^2 = 33.7$, $p < 10^{-8}$). Notably, P(IK) itself adds little to verification prediction beyond confidence and log-probability (all trials: $\chi^2 = 7.5$, $p = .006$, with no gain in cross-validated AUROC; within incorrect: $\chi^2 = 1.5$, $p = .22$), converging with the weight-vector orthogonality reported below:a model-specific estimate of question level difficulty is largely uninformative about its trial-level verification judgment.

In a nested logistic regression including P(IK), verbal confidence, verification response, and mean answer log-probability as baseline predictors, adding the PANL probe decision function significantly

improved prediction of both answer change (AUROC: .919 → .958; LR $\chi^2 = 101.7$, $p < 10^{-23}$) and A2 correctness (AUROC: .915 → .934; LR $\chi^2 = 99.0$, $p < 10^{-23}$). Surface text features (TF-IDF lexical features and token lengths of question and answer text) added no predictive value over the behavioural scalars ($p = 1.0$), confirming that the probe is not exploiting shallow textual cues.

To assess whether the PANL error-detection signal is redundant with a model-specific estimate of question-level difficulty (i.e. P(IK), we compared the weight vectors of two linear probes trained at layer 30: one predicting the verification response (V=Y vs V=N) and one predicting P(IK), a sampling-based estimate of retrieval reliability following Kadavath et al. (2022) (20 samples at temperature $= 1$; see §A.3). The cosine similarity between these weight vectors was $+0.007$, indicating near-complete orthogonality. The PANL verification probe and the P(IK) probe extract geometrically independent directions from the same activation space: one encodes the model's evaluative judgment about a specific question–answer pair, while the other encodes how reliably the model can retrieve information about the topic in general (or how difficult the question is). This dissociation confirms that the error-detection signal is not a proxy for question difficulty, and that PANL carries instance-specific evaluative information beyond what question-level knowledge estimates capture.

That the correctability signal survives control for P(IK) may seem surprising: predicting which revisions will succeed surely depends on knowledge. Indeed, knowledge explains much of correctability where it can be measured well—on PopQA, the maximal knowledge-augmented baseline approaches PANL alone (.77 vs .785). But PANL adds significantly beyond it even there, because P(IK) summarises only part of what determines correctability. P(IK) counts how often the correct answer is retrieved across samples, but is blind to the structure of the errors: at the same P(IK), the failed samples may scatter across many different wrong answers (the committed error is incidental, and revision has a good chance of reaching the correct alternative) or converge on a single systematic confusion (the same error re-emerges on revision). PANL, computed over the specific question–answer pair, can distinguish these cases; P(IK), as a scalar hit rate, cannot. This is consistent with Orgad et al. (2024), who show that error types defined by exactly this resampling structure—e.g. consistently wrong versus occasionally correct—are decodable from internal representations, and that models can internally encode the correct answer while reliably generating an incorrect one.

F.4 Error Detection: Causal Experiment Details

Activation patching: full results. Figure 5A shows d' as a function of patching layer for five token positions. Patching LAT rescues error detection at early-to-mid layers (peak $d' = 1.23$ at L15, ∼100% recovery), declining sharply after L30. Patching PANL rescues at mid layers (peak $d' = 0.89$ at L30, ∼74% recovery), with a complementary temporal profile—rising as LAT declines. Patching the prompt last token rescues at later layers (peak $d' = 1.25$ at L35, ∼107% recovery), consistent with this position assembling the final verification decision from information propagated forward through the residual stream. Patching PANL+1 and PANL offset 9—positions where linear probes successfully decode verification-predictive information—has no effect on d' at any layer, remaining at the corrupted baseline throughout. This dissociation between decodable information and causal influence is consistent with prior findings that linearly decodable signals are not always expressed in downstream behaviour (Azaria & Mitchell, 2023). The verification logprob difference ($\log P(\text{Y}) - \log P(\text{N})$) shows the same pattern (Figure A3), confirming that patching affects not only the binary verification decision but also its graded strength.

Activation noising (through mean ablation): full results. Ablating PANL alone has no detectable effect on d' at any layer (maximum $|\Delta d'| = 0.062$ at L27), with error-detection sensitivity remaining at the clean baseline ($d' = 1.20$) throughout. Ablating LAT severely disrupts error detection at early-to-mid layers ($d' = -0.14$ at L0; -0.18 at L10; -0.21 at L22), with gradual recovery from L25 ($d' = 0.44$) reaching baseline by L30 ($d' = 1.26$). Ablating the prompt last token produces no effect at early-to-mid layers but severe disruption at later layers ($d' = 0.0$ at L40–60). Ablating PANL+1 and PANL offset 9 has no effect at any layer (all $|\Delta d'| < 0.03$). The verification logprob difference shows the same pattern (Figure A3): LAT ablation produces large negative shifts (e.g., $\Delta = -20.3$ at L24) at early-to-mid layers, prompt last token ablation at later layers ($\Delta = -8.3$ at L60), and PANL ablation produces no reliable shift (maximum $|\Delta| = 0.90$).

Joint ablation details. Jointly ablating LAT and PANL produces deficits exceeding LAT ablation alone at mid layers. At L27, LAT alone yields $d' = 0.82$ (Δlogit_diff $= -15.4$) while joint ablation yields $d' = 0.31$ (Δlogit_diff $= -19.4$). At L28, LAT alone yields $d' = 1.05$ (Δlogit_diff $= -11.8$) while joint ablation yields $d' = 0.58$ (Δlogit_diff $= -17.0$). The effect is concentrated at mid layers (25–29) and dissipates by L30.

LAT confound. LAT serves a dual role: it carries both the answer representation itself and any evaluative signal computed over that content. Corrupting the answer token necessarily corrupts both, making it impossible to determine from LAT manipulations alone whether disruption reflects loss of the answer representation, loss of the evaluative signal, or both. PANL provides a cleaner window onto the evaluative computation: it is not part of the answer itself, yet carries a rich evaluative signal that is causally sufficient to restore error detection. For this reason, probing analyses focus on PANL as the primary locus of the evaluative representation, while acknowledging that the causal evidence is consistent with the evaluative signal being distributed across both positions.

F.5 Self Correction: Activation Noising

Why noising is the primary causal test for this outcome. The corrupt-and-restore patching paradigm is not a principled intervention for self-correction: the task requires the model to see the answer it committed to, but corrupting the answer span removes precisely this information, collapsing the task into unconditioned re-answering of the question. Mean-ablation noising preserves the task: the committed answer remains visible in the input, the model behaves naturally, and targeted lesions reveal which internal positions the corrective computation depends on. Note that mean-ablating LAT in the noising experiment does not reintroduce this confound: the ablation replaces a single position's residual-stream state at a single layer while all input embeddings—including the full answer span—remain intact, so the committed answer stays visible to the model through the unablated answer positions and through layers below the intervention.

Noising results (Figure A5). In Gemma 3 27B ($n = 122$; clean baseline $P(\text{TOP=GT}) = 0.66$), ablating LAT suppresses correction across all probed layers, while ablating PANL or PANL+1 alone produces no deviation from the clean baseline. Jointly ablating LAT and PANL reduces $P(\text{TOP=GT})$ below LAT-alone at mid layers—the same redundancy signature observed for verification: PANL carries a parallel corrective signal that becomes causally apparent only when LAT is also disrupted. Qwen 2.5 7B ($n = 76$; clean baseline 0.57) replicates this architecture: LAT ablation produces the dominant deficit, PANL+1 is inert, and joint LAT+PANL ablation falls below LAT-alone across mid layers. Unlike Gemma, PANL ablation alone produces a modest deficit at middle layers (L10–17), indicating that PANL carries a partially independent corrective signal in this model. In both models the corrective signal is anchored on LAT, with PANL contributing additional information, and the largest deficits occur at early-to-mid layers.

Notably, the direction of the LAT effect runs counter to a naive account in which the A2 response merely echoes the in-context A_1 token: under that account, weakening LAT should reduce in-context bias toward A_1 and *facilitate* switching to GT. Instead, LAT ablation suppresses corrective output, indicating that the LAT residual stream carries information used in the model's evaluation of its own answer rather than a literal token representation.

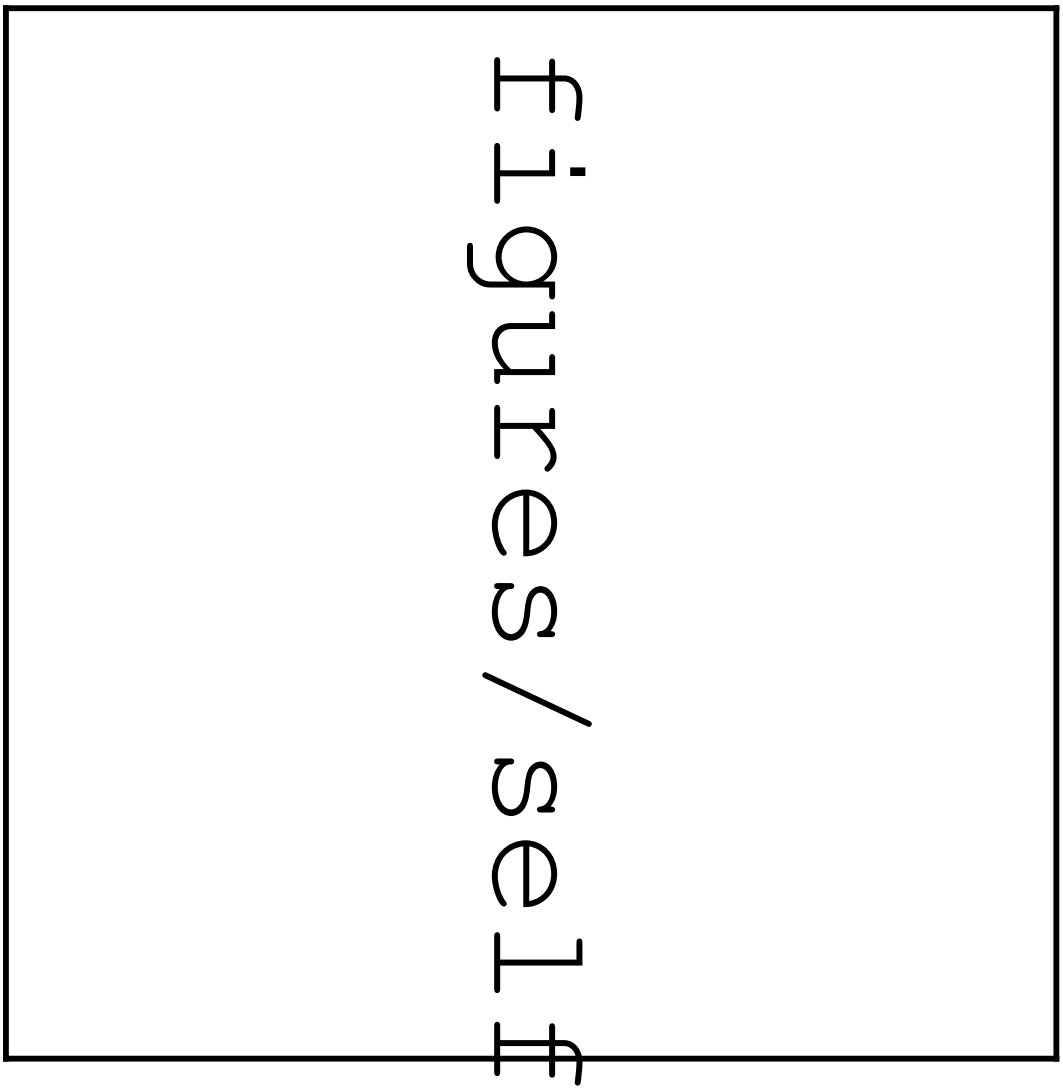

Figure A5: **Activation noising experiments relating to self-correction: Gemma 3 27B and Qwen 2.5 7B. Top panel.** Prompt. The model is shown a question, its own initial answer A_1, and a prefilled negative verification ("N"), then asked for a final answer with no explanation. LAT denotes the final token of A_1; PANL denotes the newline token immediately following A_1. The verification response is prefilled as"N" (rather than sampled) so that it is held fixed across trials and intervention conditions. This ensures that the effect of an intervention at LAT or PANL is measured on the correction step itself, rather than being confounded by the intervention also altering a sampled verification response. **Bottom panels.** Activation noising: mean-ablation applied to the residual-stream representation at a single position and layer (input embeddings intact). Analyses restricted to trials where the clean baseline elicits correction i.e. log-probability of first ground truth token is greater than that of the first A1 (initial answer) token: $n = 122$ (Gemma) and $n = 76$ (Qwen). Dependent measure in the experiment is P(top token = GT) at the answer-output position. **Gemma 3 27B.** Clean baseline (dashed) P(top=GT) = 0.66, i.e. on 66% of these trials GT is the model's top output token under the clean (unablated) prompt – for completeness, note in the rest of trials another token (not the first token of A1) is the top token. Ablation of the last answer token (LAT) suppresses correction across all layers; PANL and PANL+1 ablation alone produce no deviation from the clean baseline. Joint LAT+PANL ablation reduces P(top=GT) below LAT-alone at mid layers. The direction of the LAT effect runs counter to a naive in-context-echo account (under which weakening LAT should facilitate switching to GT), indicating LAT carries evaluative rather than literal token information. **Qwen 2.5 7B.** The same pattern replicates: LAT ablation produces the dominant deficit, PANL+1 is inert, and joint LAT+PANL falls below LAT alone across mid-layers. Unlike in Gemma, PANL ablation alone (dark blue) produces a modest deficit in the middle layers (L10–17, dipping below the clean baseline), indicating that PANL carries a partially independent corrective signal in this model. In both models the corrective signal is anchored on LAT, with PANL contributing additional information; the largest deficits occur at early-to-mid layers. Qwen clean baseline P(top=GT) = 0.57.

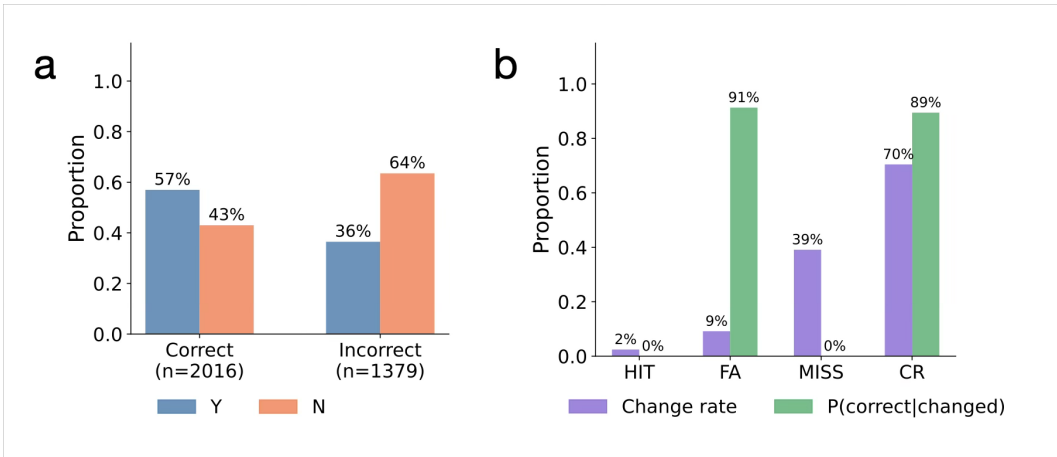

Figure A6: MNLI neutral behavioural results (n = 3,395). **a**: Verification responses by A1 correctness (d' = 0.53, c = 0.15). Unlike TriviaQA, the model shows a near-neutral criterion. **b**: Answer change rate and correction success by SDT cell. Verification gates revision as in TriviaQA, but correction success is concentrated in correct rejections (89.6%); misses yield 0% correction by construction (changing away from the correct label is always wrong in 3-way classification).

G Cross-Model, Cross-Task Generalisation and Practical Relevance

G.1 MNLI Behavioural Results

Gemma 3 27B scored 71.6% on the MNLI dataset at first attempt (A1: n = 9,888). Self-correction yielded a modest but significant improvement to 73.9% at A2 (Δ = +2.3 percentage points; McNemar's $\chi^2 = 50.6$, $p < .001$) — though there was a marked improvement within neutral trials (see below). Across all labels, the model displayed weak error-detection sensitivity ($d' = 0.70$).

Answer change rates varied dramatically by label: the model changed its answer on only 0.8% of entailment trials and 3.3% of contradiction trials, compared to 30.3% of neutral trials. We therefore restrict all subsequent MNLI analyses to neutral trials (n = 3,395), where the model displays meaningful self-correction behaviour.

Verification behaviour on neutral trials. Figure A6 shows the distribution of verification responses as a function of A1 correctness for neutral trials. The model detects its errors at a modest rate (d' = 0.53, c = 0.15), with a near-neutral criterion—markedly different from the strong Y-bias observed in TriviaQA ($c = -1.34$). When A1 is correct, the model confirms its answer 54.7% of the time; when A1 is incorrect, the model responds V=Y on 33.9% of trials. Self-correction on neutral trials yields a marked improvement in accuracy from 59.4% to 66.1% (Δ = +6.7 percentage points; McNemar's $\chi^2 = 53.1$, $p < .001$).

Verification determines whether the model revises, but not whether revision succeeds. Figure A6 shows answer change rates and correction success by SDT cell for neutral trials. As in TriviaQA, the verification response gates revision: correct rejections show a 72.7% change rate compared to 0.2% in false alarms, and misses show 40.1% compared to 0.1% in hits. Conditional on changing, correction success is high in correct rejections (89.6%) but by nature near zero in misses (0.0%), a pattern that differs from TriviaQA's roughly constant ∼30–34% correction rate across cells. This difference likely reflects the structure of the task: in MNLI, switching from neutral to the correct alternative (entailment or contradiction) requires identifying a specific label, and the model appears substantially better at this when it has correctly detected an error (CR).

Confidence drives the verification decision but is a weaker signal than in TriviaQA. Verbal confidence carries evaluative information beyond log-probabilities on neutral trials, replicating the

second-order behavioural signature observed in TriviaQA. Adding verbal confidence to a logistic model containing only the mean answer log-probability improves prediction of the verification response (LR $\chi^2 = 13.5$, $p < .001$; $\beta_{\text{conf}} = +0.069$, $\beta_{\text{logprob}} = +0.003$). Confidence also adds beyond log-probability and A1 correctness combined (LR $\chi^2 = 17.2$, $p < .001$). Within incorrect trials, confidence predicts verification (AUROC $= .567$, $\chi^2 = 9.3$, $p = .002$). However, these effects are substantially weaker than in TriviaQA ($\beta_{\text{conf}} = 0.069$ vs 0.51; within-incorrect AUROC $= .567$ vs .737), suggesting that the model's evaluative signal for NLI is less robust than for factual QA.

Verbal confidence correlates weakly with the verification logprob difference ($r = .090$, $p < .001$; within correct $r = .086$; within incorrect $r = .145$), a much weaker association than in TriviaQA ($r = .604$). Mean answer log-probability is essentially unrelated to verification ($r = -.022$, $p = .20$), confirming that generation-time fluency provides no useful signal for NLI verification. Unlike TriviaQA, where verbal confidence followed a clean SDT-consistent ordering (HIT $>$ FA $>$ MISS $>$ CR), the pattern in MNLI neutral shows: FA (0.614) $>$ HIT (0.566). The reversal of HIT and FA is notable: the model expresses higher confidence on incorrect trials it endorses than on correct trials it endorses, the opposite of TriviaQA. The compressed overall confidence range further indicates that the model's confidence signal carries less information about its own accuracy on this task.

Verbal confidence or verification confidence does not predict correction quality. As in TriviaQA, verbal confidence does not predict which revisions succeed. Within all changed trials, confidence is uninformative about A2 correctness (AUROC $= .529$, $r = +.056$, $p = .072$, $n = 1{,}030$). Within incorrect changed trials, the same null holds (AUROC $= .527$, $r = -.035$, $p = .37$, $n = 663$). The verification logprob difference similarly fails to predict correction success within changed trials (AUROC $= .565$, $r = -.115$, $p < .001$, $n = 1{,}030$; within incorrect changed: AUROC $= .552$, $r = -.058$, $p = .14$, $n = 663$). Both signals determine whether the model revises but not whether revision succeeds—replicating the TriviaQA dissociation and motivating the PANL probing analyses that follow.

G.2 MNLI: PANL Activations Predict Error Detection and Self-Correction

We asked whether the PANL evaluative signal identified in TriviaQA generalises to a structurally different task. MNLI neutral trials present a distinct challenge: the model's behavioural confidence is weakly informative (AUROC $= .534$ for verification prediction), the confidence range is compressed, and the SDT-cell ordering does not follow the clean pattern observed in TriviaQA. If PANL encodes a domain-general evaluative signal, it should nonetheless predict verification and correction outcomes beyond behavioural baselines.

PANL predicts verification responses beyond behavioural signals. PANL activations at layer 30 predict verification responses with AUROC $= .868$, far exceeding verbal confidence alone (AUROC $= .534$) and the control position (third question token: AUROC $= .503$). Within incorrect trials, PANL achieves AUROC $= .813$ (control: .499). The gap between PANL and behavioural baselines is even larger here than in TriviaQA, because the behavioural signals are so weak on this task—the model's overt confidence captures very little about its own verification behaviour on MNLI neutral, yet the internal representation at PANL carries a strong signal.

PANL predicts answer changes. PANL at layer 30 predicts whether the model will change its answer with AUROC $= .869$, exceeding the combined behavioural baseline of verbal confidence, A1 correctness, and verification response ($\Delta = +.045$ above baseline AUROC $= .823$). The control position shows no improvement (AUROC $= .523$). Within incorrect trials, PANL achieves AUROC $= .811$ ($\Delta = +.024$ above baseline .787).

PANL predicts self-correction success where behavioural signals fail. Replicating the central TriviaQA finding, PANL predicts which revisions will succeed where behavioural signals cannot. Across all trials, PANL predicts A2 correctness with AUROC $= .697$ ($\Delta = +.143$ above baseline .555; control: .541). The effect is strongest in the subsets where behavioural signals are at chance. Within trials where the model changed its answer—where verbal confidence is uninformative about correction success (AUROC $= .529$)—PANL achieves AUROC $= .857$ ($\Delta = +.335$ above baseline .522). Within incorrect trials that changed—the subset where self-correction matters most, and

where the combined behavioural baseline is at chance (.512)—PANL achieves AUROC = .862 ($\Delta = +.351$). The control position shows no predictive power in any of these subsets ($\leq .543$).

The correctability prediction is numerically stronger on MNLI neutral than on TriviaQA (AUROC = .862 vs .614 within incorrect trials), likely reflecting the higher base rate of successful correction on this task (89.6% in CR trials vs ∼34% in TriviaQA). Nonetheless, the qualitative pattern replicates: PANL encodes a graded signal about whether the model can produce a correct second-pass answer, and this signal is not reducible to any combination of behavioural confidence measures.

Summary. The PANL evaluative signal generalises from factual QA to natural language inference. Despite substantially weaker behavioural signals, PANL activations at layer 30 predict verification, answer change, and—most critically—correction success on MNLI neutral trials, replicating all key findings from TriviaQA. The model's internal evaluative representation is richer than what its behavioural signals convey, across both tasks.

G.3 Pop QA: Behavioural Results

PopQA (Mallen et al., 2023) comprises 13,068 entity-centric factual questions derived from Wikidata, spanning 16 relation types (e.g. director, screenwriter, capital, country). Questions are templated ("Who was the director of X?") with short entity answers (mean 1.9 words), providing a structurally distinct complement to TriviaQA's diverse, naturally-phrased questions. On PopQA, calibrated answer logprobs achieve AUROC = .816 and ECE = .205 for predicting A1 correctness, while verbal confidence achieves AUROC = .764 and ECE = .399. Verification results showed robust error-detection capacity (Figure A7a), and verbal confidence tracked A1 correctness within each verification cell (Figure A7b). As in TriviaQA, verification gates whether the model revises, but correction success is roughly constant (at 10% across all categories; excluding the Hits cell where there are only 2% revisions; Figure A7). Self-correction improved accuracy significantly from 29.3% to 31.7% ($\Delta = +2.4\%$, McNemar's $\chi^2 = 197.3$, $p < 10^{-44}$, $n = 13{,}068$). Self-correction success (average 10%) was strongly correlated with domain knowledge (max 40% for capital cities, where A1 correctness 80%), as indexed by average A1 accuracy within a question category (Figure A7).

On PopQA, the same pattern for verification prediction holds as for TriviaQA (Figure A7d). Logprobs alone achieve AUROC = .616 for predicting the verification response, while verbal confidence alone achieves .776; adding confidence to logprobs produces a large improvement (LR $\chi^2 = 2007.6$, $p \approx 0$), with confidence dominating ($\beta_{\text{conf}} = +0.71$, $\beta_{\text{logprob}} = +0.03$). Within incorrect trials, confidence remains strongly predictive (AUROC = .737—identical to TriviaQA) while logprobs are near chance (.532). Verbal confidence thus carries evaluative information beyond generation-time logprobs on this dataset as well, consistent with a second-order signal that generalises across factual QA formats.

G.4 PopQA: PANL Activations Predict Error Detection and Self-Correction

On PopQA, PANL activations again substantially exceed behavioural baselines for predicting the verification response (Figure A7d,f). At layer 30, PANL achieves AUROC = .993 against a combined behavioural baseline of .801; within incorrect trials, PANL AUROC = .991 versus .732. The within-incorrect separation confirms that PANL discriminates detected from undetected errors on this dataset, with performance comparable to TriviaQA (.991 vs .958) despite the different question format and lower base accuracy (29.3% vs 75.5%). For predicting A2 correctness within incorrect trials that changed answer ($n = 3{,}688$), the behavioural baseline is substantially higher on PopQA than on TriviaQA (.671 vs .475). Critically, PANL still adds above the full behavioural baseline (.785 vs .671, $\Delta = +.114$), confirming that the evaluative signal at PANL carries information beyond behavioral measures.

PANL encodes a trial-specific evaluative signal, rather than question difficulty or frequency in the training data We considered the possibility that PANL's predictive power in the PopQA dataset reflects surface-level properties of the question rather than a genuine evaluative signal—for instance,

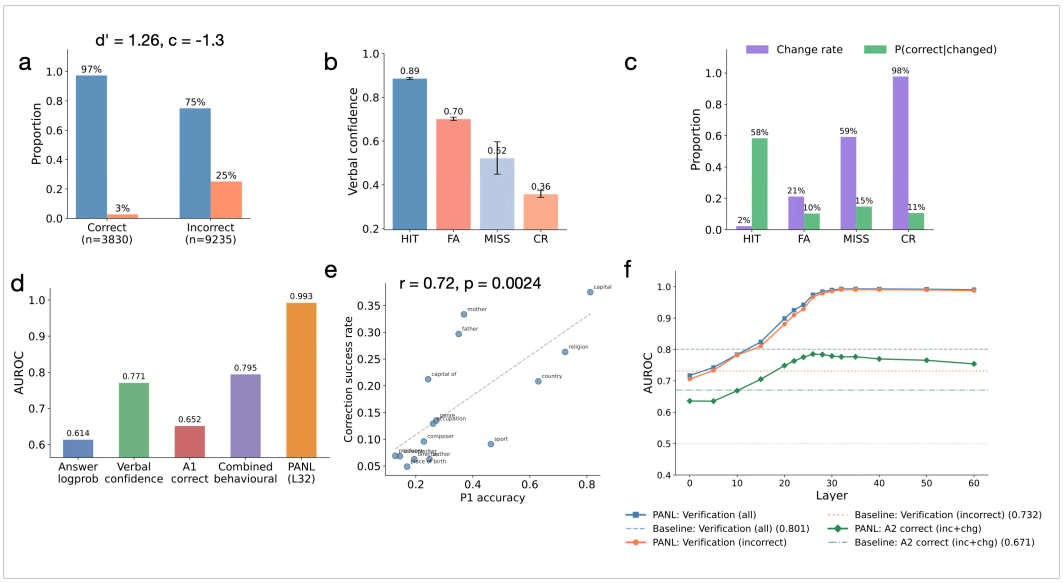

Figure A7: **Gemma 3 27B behavioural results on PopQA dataset**. 13,068 entity-centric factual questions derived from Wikidata, spanning 16 relation types (e.g. director, screenwriter, capital, country)(Mallen et al., 2023). **a**: Verification responses by A1 correctness. The model displays robust error detection ($d' = 1.26$) with a strong bias to respond Yes ($c = -1.30$) **b**: Mean verbal confidence ($\pm$95% CI) across 4 cells (i.e. Hits/False Alarms/Misses/Correct Rejections; e.g. Hit = A1-correct given a Y response at verification (i.e. V=Y). FA: A1-incorrect given V=Y). **c**: Answer change rate (purple) (i.e. proportion of answers changed between A1 and A2) – and correction success conditional on changing (green) by each of the 4 cells (i.e. Hits/False alarms etc). Note: Self-correction improved accuracy significantly from 29.3% to 31.7% ($\Delta = +2.4\%$, McNemar's $\chi^2 = 197.3$, $p < 10^{-44}$, $n = 13{,}068$). **d**: AUROC for predicting the verification response from answer token log-probabilities, verbal confidence, A1 correctness, their combination and PANL activation at layer 30. Verbal confidence substantially outperforms log-probabilities, consistent with a second-order evaluative signal that is partially independent of generation fluency. PANL outperforms best behavioural baseline (all signals combined). **e**: Correlation between correction success and A1 correctness within a PopQA category. Higher correction success where there is more knowledge of that domain. **f**: PANL probe AUROC across layers for predicting verification (all trials, blue), verification within incorrect trials (orange), and A2 correctness within incorrect changed trials (green). Dashed lines indicate combined behavioural baselines. PANL exceeds all three baselines at mid-to-upper layers, peaking around L30. The advantage persists after augmenting the baseline with Wikipedia page-view frequency (`s_pop`, a proxy for entity popularity and training-data exposure) and P(IK) (a sampling-based index of retrieval reliability or question difficulty (Kadavath et al., 2022)): PANL still adds above this maximal first-order baseline for A2 correctness ($\chi^2 = 104.3$, $p < 10^{-24}$). For verification, `s_pop` alone is uninformative (AUROC = .502). These additional results confirm that PANL captures a trial-specific evaluative signal rather than question difficulty or entity popularity.

questions about popular entities may be both easier to verify and easier to correct. Hence PANL might encode something akin to entity popularity. To test this, we regressed A2 correctness (within incorrect+changed trials) on Wikipedia page-view frequency (`s_pop`, a proxy for training-data exposure; Mallen et al., 2023). We also considered the possibility that PANL might encode question difficulty, or the model's ease of retrievability of the answer – captured by the P(IK) index: the fraction of 20 temperature-sampled answers (temp = 1.0) matching the modal response (Kadavath et al., 2022).

The maximal first-order baseline including `s_pop`, P(IK), logprob, verbal confidence, and verification logprob difference reaches .77. PANL adds significantly above this maximal baseline (i.e. .81; $\chi^2 = 104.3$, $p < 10^{-24}$, ΔAUROC $= +.035$) – and PANL alone (.78) exceeds this strong baseline. For verification prediction, the contrast is even starker: PANL alone achieves .993 while `s_pop` is uninformative (.502) and the maximal behavioural+`s_pop`+P(IK) baseline reaches only .816.

These findings are consistent with PANL capturing a trial-specific evaluative signal—whether this particular answer fits this particular question—that is partially independent of how well-known the entity is, a proxy for question difficulty or how consistently the model retrieves it (i.e. P(IK)).

G.5 Cross-Task Transfer

To test whether the evaluative signal generalises across datasets, we trained PANL L30 probes on one dataset and applied them to the other (Table A8). Verification probes transfer near-perfectly between TriviaQA and PopQA in both directions (T→P: .982; P→T: .988). This contrast with MNLI is informative: TriviaQA and PopQA differ in question structure (diverse vs. templated) but share the same task domain (factual QA), whereas TriviaQA and MNLI differ in both domain (factual QA vs. natural language inference) and answer format (free-form vs. three-way classification). TriviaQA–MNLI verification transfer is asymmetric (T→M: .587; M→T: .934): the MNLI-trained direction captures a general evaluative feature that is also present in TriviaQA's richer representational space, while the TriviaQA-trained direction is too task-specific to generalise. The near-perfect bidirectional transfer between factual QA datasets, together with the higher cosine similarity between their verification directions (.397 vs .132 for TriviaQA–MNLI), suggests that PANL encodes a domain-general evaluative signal within factual QA that partially overlaps with—but is not identical to—the evaluative signal for NLI.

Correctability transfer is more moderate (T→P: .663; P→T: .610), but exceeds TriviaQA–MNLI correctability transfer (T→M: .547; M→T: .477, at chance). Cosine similarity between correctability directions is low for both dataset pairs ($\leq$ .071). This suggests that while the *existence* of a correctability signal at PANL is universal, the specific representational direction encoding it is more task- and dataset-specific than the verification direction—consistent with correctability reflecting task-specific knowledge structures.

G.6 Practical relevance

To test whether the PANL evaluative signal is practically useful, we used it to selectively route trials to self-correction. A 5-fold cross-validated logistic regression probe is trained on PANL activations at layer 30 to predict the deployment-relevant target "switch helps" (A_1 wrong AND A_2 correct); ground truth and A_2 outcomes are used only to construct training labels. At deployment, the probe scores each trial from its PANL activation alone, at the end of the first-pass answer, and trials scoring above a threshold are routed to the self-correction prompt (see Table A9 caption for full details).

The probe achieves AUROC $= 0.90$ (TriviaQA) and 0.86 (PopQA) for predicting which trials would benefit from correction. Routing the top 50% of trials by probe score recovers the accuracy of unconditional self-correction within 0.1 percentage points on both datasets (TriviaQA: 79.2% vs 79.3%; PopQA: 31.6% vs 31.7%), while invoking the correction prompt on only half of all trials—a $2\times$ reduction in correction-time inference cost with negligible accuracy loss. At the top 10% threshold, the probe achieves accuracy comparable to routing on the model's own verification token

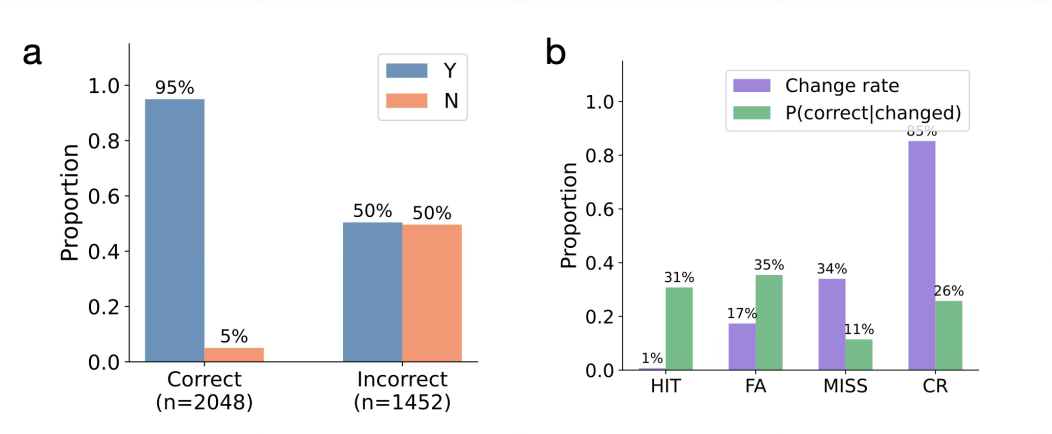

Figure A8: Qwen 2.5 7B verification and self-correction results on TriviaQA ($n = 3{,}500$). **a**: Verification responses by A1 correctness ($d' = 1.63$, $c = -0.83$). **b**: Answer change rate (purple) and correction success conditional on changing (green) by SDT cell. The pattern replicates Gemma 3 27B: verification gates revision, but correction quality is roughly constant across cells.

(V=N) at similar selectivity, while additionally offering a tunable cost–accuracy trade-off that the binary verification signal does not provide. A simple linear probe over PANL activations thus enables selective self-correction with substantial inference savings.

G.7 Qwen 2.5 7B Behavioural Results

Qwen 2.5 7B scored 58.5% on the TriviaQA dataset at first attempt (A1: $n = 3{,}500$). Self-correction yielded a significant improvement to 63.2% at A2 ($\Delta = +4.7$ percentage points; McNemar's $\chi^2 = 106.3$, $p < .001$), a larger gain than observed in Gemma 3 27B (+3.7 pp).

Verification behaviour. The model displays robust error-detection sensitivity ($d' = 1.63$, $c = -0.83$), comparable to Gemma ($d' = 1.67$, $c = -1.34$) though with a less extreme Y-bias (see Figure A8). When A1 is correct, the model confirms its answer 95.0% of the time; when A1 is incorrect, the model responds V=Y on 50.4% of trials, detecting its error on approximately half of incorrect trials — a substantially higher detection rate than Gemma (30.8%).

Verbal confidence carries predictive information beyond log-probabilities. As in Gemma, verbal confidence alone predicts verification (with AUROC = .703); answer log-probability alone achieves AUROC = .750; the combined model yields AUROC = .812. Within incorrect trials, verbal confidence achieves AUROC = .655 while answer log-probability drops to AUROC = .534, consistent with the pattern observed in Gemma. The correlation between verbal confidence and the verification logprob difference is $r = .535$ (within correct $r = .371$; within incorrect $r = .457$).

Verification determines whether the model revises, but not whether revision succeeds. The verification response gates revision as in Gemma: correct rejections show an 85.3% change rate compared to 17.3% in false alarms, and misses show 34.0% compared to 0.7% in hits (see Figure A8).

As in Gemma, conditional on changing, correction quality is roughly constant across cells: 30.8% in hits ($n = 13$, interpret cautiously), 35.4% in false alarms, 11.4% in misses, and 25.7% in correct rejections. The overall correction rate among changed incorrect trials is 27.4% (see Figure A8).

Neither confidence signal predicts correction quality. As in Gemma, neither verbal confidence nor the verification logprob difference predicts which revisions succeed. Within incorrect changed trials, verbal confidence is uninformative about A2 correctness (AUROC = .528, $r = +.042$, $p = .25$, $n = 741$). The verification logprob difference shows the same null (AUROC = .506, $r = -.019$, $p = .61$). Both signals determine whether the model revises but not whether that revision succeeds, replicating the dissociation observed in Gemma and motivating the PANL probing analyses.

Summary. Qwen 2.5 7B replicates the key behavioural findings from Gemma 3 27B despite being a substantially smaller model (7B vs 27B) with lower baseline accuracy (58.5% vs 75.5%). Error detection is robust ($d' = 1.63$), verbal confidence carries evaluative information beyond log-probabilities, and confidence predicts whether errors are detected but not whether they can be corrected. The self-correction improvement is slightly larger than Gemma ($+4.7$ vs $+3.7$ pp), and the model detects a higher proportion of its errors (49.6% vs 30.8%), though with a lower correction success rate when it does change (27.4% vs $\sim$34%).

G.8 Qwen 2.5 7B: PANL Activations Predict Error Detection and Self-Correction

We tested whether the evaluative signal replicates in Qwen 2.5 7B, a smaller model with 28 layers. Figure A9 and Table A7 summarise the probing results.

The key findings replicate: PANL predicts verification with AUROC $= .961$ ($\Delta = +.101$ above the behavioural baseline), and within errors only achieves .917 where the behavioural baseline drops to .646. Critically, PANL predicts correction success within incorrect trials (AUROC $= .774$, $\Delta = +.055$) and within incorrect changed trials (.679, $\Delta = +.104$), where behavioural predictors are near chance (Table A7). PANL also tracks the graded strength of the verification decision ($r = .909$ with the verification logprob difference). The evaluative representation at PANL thus generalises across model families and scales, replicating in a 7B model despite lower baseline accuracy (58.5% vs 75.5%) and a shallower layer hierarchy.

G.9 Qwen 2.5 7B: Causal Experiments

We replicated the causal experiments on Qwen 2.5 7B (Figure A10). Corruption of answer tokens reduced d' from 1.44 to 0.83 — a weaker disruption than in Gemma ($d' = 0.09$), likely because question tokens alone carry sufficient difficulty information to support partial error detection in this model.

Activation patching reveals the same three-position pattern: LAT rescues at early layers (peak $d' = 1.19$ at L10, $\sim$59% recovery), PANL at mid layers (peak $d' = 1.25$ at L15, $\sim$70% recovery), and the prompt last token at later layers (peak $d' = 1.45$ at L19, full recovery). PANL+1 shows no rescue at any layer. PANL is thus causally sufficient for error detection in Qwen, as in Gemma.

Mean ablation confirms the complementary pattern: ablating LAT severely disrupts error detection at early-to-mid layers ($d' = -0.30$ at L10), with recovery by L16. Ablating PANL alone has no effect ($|\Delta d'| \leq 0.14$). Jointly ablating LAT and PANL produces an additional deficit at the critical transition layers where LAT is recovering: at L15, LAT alone yields $d' = 0.93$ while the joint ablation yields $d' = 0.45$. As in Gemma, the additional PANL deficit is visible only when LAT is also disrupted, consistent with the two positions carrying the evaluative signal redundantly.

The causal architecture thus replicates across model families: PANL is sufficient but not necessary, LAT is necessary at early layers, and both positions carry the evaluative signal redundantly at mid layers.

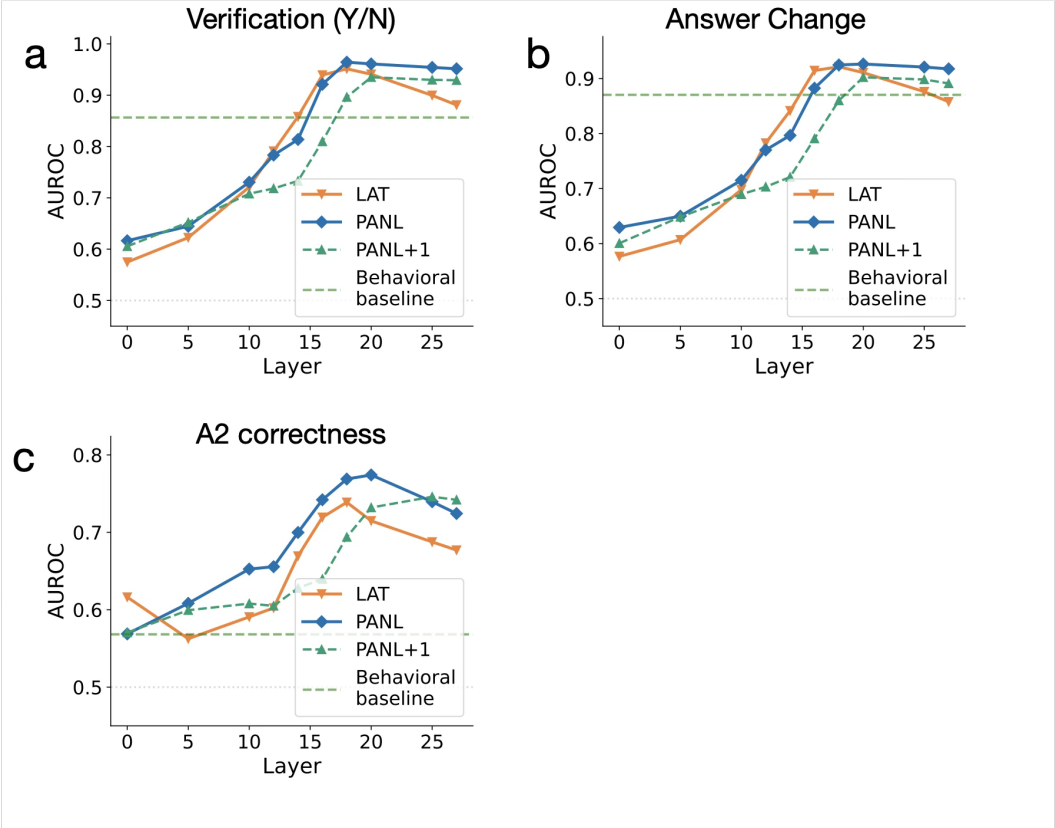

Figure A9: Qwen 2.5 7B linear probe results on TriviaQA ($n = 3{,}500$). AUROC for predicting verification response, answer change, and A2 correctness from residual stream activations at three token positions: LAT (orange), PANL (blue), and PANL+1 (green). Dashed lines: behavioural baseline (A1 correctness + mean answer log-probability + verbal confidence). **a**: Verification response (V=Y vs V=N). **b**: Answer change. **c**: A2 correctness within A1 incorrect trials. All three targets show probing performance exceeding the behavioural baseline at mid-to-upper layers, with LAT and PANL performing comparably — replicating the Gemma 3 27B pattern in a smaller model with a shallower layer hierarchy (28 layers vs 62).

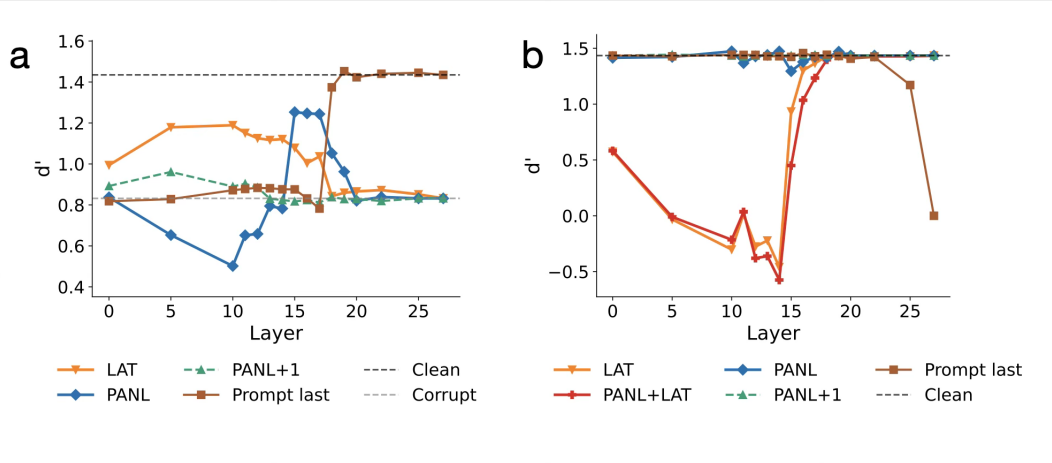

Figure A10: Qwen 2.5 7B causal experiments on TriviaQA. **a**: Activation patching. Clean $d' = 1.44$; corrupted $d' = 0.83$ (grey dashed; the above-chance corrupt baseline likely reflects question-difficulty information carried by the question tokens, which remain intact). LAT (orange) rescues at early layers (peak $d' = 1.19$ at L10, $\sim$59% recovery); PANL (blue) rescues at mid layers (peak $d' = 1.25$ at L15, $\sim$70% recovery); prompt last token (brown) rescues at later layers (peak $d' = 1.45$ at L19, $\sim$103% recovery). PANL+1 (green dashed) shows no rescue. **b**: Activation noising experiment with mean ablation. Ablating LAT (orange) severely disrupts error detection at early-to-mid layers ($d' = -0.30$ at L10), recovering by L16. Ablating PANL alone (blue) has no effect (all $|\Delta d'| \leq 0.14$). Jointly ablating LAT and PANL (red) produces additional deficits beyond LAT alone at the critical mid-layer transition: at L15, LAT alone yields $d' = 0.93$ while the joint ablation yields $d' = 0.45$ ($\Delta = -0.48$), consistent with representational redundancy.