How Much Does a Reasoning Summary Reveal? An Observability Ladder for Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

A model's reasoning trace can expose errors and misbehavior absent from its final response, yet closed-source deployments may show users only that response and a short reasoning summary. Here, we ask how much monitoring performance this reduced access preserves and introduce an observability ladder that holds completed runs fixed and varies access to prompts, responses, reasoning summaries, full traces, and internal model features. We use final-answer correctness as an externally labeled test case across five open-weight models and three benchmarks, with each model writing a self-summary from its completed trace alone. Linear classifiers trained on text features and embeddings reach mean AUROC 0.73 from the prompt alone and 0.75 with the response. The summary then adds 0.02 (95% question-bootstrap CI $[0.01, 0.03]$), the full trace a further 0.04 $[0.03, 0.05]$, and internal features 0.03 $[0.02, 0.04]$, reaching 0.84. However, a summary-level reader given the trace's word and sentence counts without its text trails the same reader given the full text by 0.01 $[0.01, 0.02]$, so part of the trace's gain is length. Without the prompt, the summary's gain jumps to 0.16 $[0.13, 0.18]$, consistent with the summary standing in for the missing question, and for the linear classifiers, telling a correct from an incorrect run of the same MMLU-Pro question barely rises above chance (0.53 to 0.57). Much of the across-question performance, then, is attainable from the prompt alone, without observing the particular run. Two LLM readers recover far more within-question discrimination from summaries and traces than the linear classifiers when the prompt is withheld, yet reach their highest concordance from the prompt and response alone when it is visible, so what a summary reveals depends on the reader. We have measured one property and one kind of summary, but because the ladder holds the run fixed and varies only what the reader sees, it can be applied to misbehavior with independent labels, to provider summaries, and to stronger readers.

1 Introduction

Reasoning traces can expose errors or misbehavior absent from a model's final response (Korbak et al., 2025; Baker et al., 2025a; Guan et al., 2026). Yet interfaces differ in what they reveal. In closed-source deployments such as those from OpenAI, Google, and Anthropic, a user may receive only the final response or a short provider summary (OpenAI, 2024; 2026; Google, 2026; Anthropic, 2026), while the full trace, answer probabilities, and hidden states require deeper access (Baker et al., 2025b; OpenAI, 2025a; Guo et al., 2017; Kadavath et al., 2022; Burns et al., 2023; Azaria & Mitchell, 2023). Evidence from a full trace therefore does not establish what a user can learn from a response or summary, and what a display reveals depends on who reads it as well as on what it contains.

We ask how much of the monitoring performance available from the full trace survives when only a reasoning summary is shown. We evaluate final-answer correctness, whose labels are defined independently of the monitoring procedure and remain fixed across access levels. We ask three questions. First, how much does the full trace add once the prompt, response, and summary are available, and what do internal features add beyond that (Section 3)? Second, how much of any gain depends on trace length, summary construction, and whether the reader sees the prompt (Section 4)? Third, can either text distinguish correct from incorrect runs of the same question (Section 5)?

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.

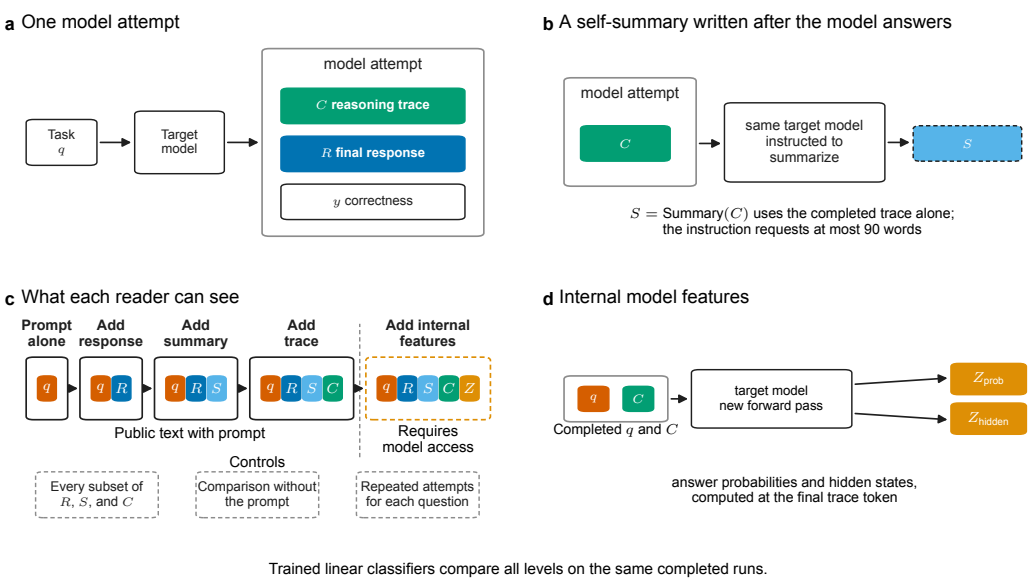


Figure 1: **The observability ladder holds a completed run fixed and expands access to its evidence. a**, A run supplies prompt q, trace C, response R, and external correctness label y. **b**, The same producer summarizes C after answering. **c**, Starting from the prompt alone, each level adds evidence while retaining the preceding sources. Prompt-withheld and repeated-question comparisons test the contribution of question context. Trained linear classifiers read every display on the same completed runs. **d**, A new forward pass through q and completed C supplies internal features.

Each target model first completes a run, and we then build different displays of that same run and evaluate each on identical held-out cases (Figure 1). The same model writes a post-hoc summary from its trace alone. This gives us a controlled proxy for a summary interface, whose construction we can inspect. We call the procedure that turns a display into a judgment a reader. Our aim is not the best possible correctness predictor but a fair comparison across access levels, so we keep the readers simple and hold the classifier family, feature-extraction method, and fitting procedure fixed across levels. Our claims therefore concern the tested readers rather than the text alone. Each level retains the preceding sources, and the component-block control also retains their representations.

1.1 Related Work

Monitorability evaluations study what reasoning reveals about misbehavior and how this depends on the monitor and its inference effort (Baker et al., 2025a; Guan et al., 2026; Emmons et al., 2025; Wang et al., 2026). In sabotage-detection settings, Arike et al. (2026) show that giving LLM monitors more of an agent's reasoning and actions can reduce performance. Xiao & Phuong (2026) find that summarizing an intermediate monitor's reasoning reduces attack detection under collusion. We examine correctness monitoring across prompt conditions, trace-count controls, and repeated runs. This complements answer verification and process supervision, which use intermediate reasoning to improve or check answers (Cobbe et al., 2021; Lightman et al., 2024). Predicting correctness does not establish that the displayed reasoning faithfully reports the computation that produced the answer (Lanham et al., 2023; Lyu et al., 2023; Turpin et al., 2023; Chen et al., 2025).

Palod et al. (2026) compare traces, summaries, and post-hoc explanations in a human study of answer-correctness judgments. We examine automated readers across controlled access levels, prompt conditions, and repeated runs of the same question. Related work also studies whether weaker models can continue a solution from stronger models' traces (Roytburg et al., 2026), or reconstruct reasoning and transfer capabilities from summaries (Zhang et al., 2026). Training-time length penalties change the reasoning process itself (Little, 2026), whereas our summaries and extracts compress the display after generation. Confidence and hidden-state probes offer further sources of diagnostic evidence (Kadavath et al., 2022; Burns et al., 2023; Azaria & Mitchell, 2023; Ghasemabadi

Table 1: **Completed runs and labels are held fixed across access levels.** Appendix Tables A11–A19 give reader token use, detailed specifications, and correctness prevalence. Figures A5–A7 show a worked run.

Component	Specification
Producer runs	Five Qwen3/gpt-oss models and three benchmarks, with 82,055 runs, one per model and item. Thinking enabled, temperature/top-p 1, limits 28,000 tokens (Qwen3) and 126,000 (gpt-oss). Multiple-choice prompts request one option letter. Mathematics requests a final answer.
Summaries	Same producer, trace alone, preservation instruction, sampled decoding, 256-token limit. The instruction asks for at most 90 words. A separate hard-capped control is included.
Labels	Option match for GPQA/MMLU and GPT-5-mini equivalence judgments for Omni-MATH, following the benchmark procedure. A run whose option cannot be extracted counts as incorrect.
Readers	Logistic classifiers over text features and 512-dimensional embeddings, with validation-selected regularization. In the appendix sensitivity, GPT-5-mini-2025-08-07 and Qwen3.5-27B return correctness probabilities. Their prompts and inference settings are in Appendix A.3.
Splits	Five correctness-stratified 60/20/20 item splits. Preprocessing uses training data, regularization uses validation, and final fitting uses train plus validation. Repeated runs inherit question grouping. Compared displays use identical held-out cases.
Populations	Held-out first runs, discordant three-run MMLU questions for within-question comparisons, and the full held-out question union for the appendix selection analysis.
Uncertainty	Paired question bootstrap, preserving shared questions, overlapping splits, displays and repeated runs. Intervals for held-out first runs use 2,000 draws, repeated-run intervals 5,000. Fitted scores are held fixed.

& Niu, 2025), as do verbalized uncertainty (Lin et al., 2022) and semantic entropy over repeated samples (Kuhn et al., 2023; Farquhar et al., 2024).

Our contribution is a controlled test of how much of the correctness-monitoring performance of full traces these model-written summaries recover, with prompt visibility, length controls, and repeated questions showing how much the comparison depends on context and population. Two LLM readers test sensitivity to the reading procedure through paired display comparisons on the same completed runs (Section 5 and Appendix A.4).

2 THE OBSERVABILITY LADDER AND EXPERIMENTAL DESIGN

We evaluate five open-weight models, Qwen3-4B, Qwen3-8B, Qwen3-14B, gpt-oss-20b, and gpt-oss-120b (Qwen Team, 2025; Yang et al., 2025; OpenAI, 2025b; Agarwal et al., 2025), on GPQA-Diamond, MMLU-Pro, and Omni-MATH-2-Filtered (Rein et al., 2024; Wang et al., 2024; Ballon et al., 2026). The benchmarks contain 198, 12,032, and 4,181 items, respectively, giving 82,055 runs with one run per model and item. We call the model that produced a run its producer. Table 1 summarizes the experiment.

For each completed run, we record prompt q, response R, self-summary S, full trace C, internal features Z, and correctness y. We obtain Z by replaying the prompt and completed trace through the target model in a new forward pass, extracting the last-token hidden state and answer probabilities. The prompt-visible ladder, starting from the prompt alone, is

$$\begin{aligned} &O_0 = (q), && O_1 = (q, R), && O_2 = (q, R, S), \\ &O_3 = (q, R, S, C), && O_4 = (q, R, S, C, Z). \end{aligned} \tag{1}$$

The corresponding prompt-withheld ladder omits q, although internal features always see q and C. We call displays without Z public, since they need no model access. We separately evaluate all public component subsets as controls for the order in which sources are added.

For a specified reader r and performance measure A_r, the ordered increments are

$$\begin{aligned} \Delta_S^{(r)} &= A_r(O_2) - A_r(O_1), \\ \Delta_C^{(r)} &= A_r(O_3) - A_r(O_2), \\ \Delta_Z^{(r)} &= A_r(O_4) - A_r(O_3). \end{aligned} \tag{2}$$

These measure summary added after response, trace added after summary, and internal features added after the full public display. The secondary contrast

$$\Delta^{(r)}_{C-S} = A_r(q, R, C) - A_r(q, R, S) \tag{3}$$

measures trace substituted for summary. It answers a different question from retaining the summary and adding the trace.

We specify each reader by its representation, fitting procedure, and inference method. The primary linear classifier uses rate-based text features (surface features such as length, digit density, and per-word rates of uncertainty markers) and a 512-dimensional Cohere embed-v4.0 embedding of the entire display, with no truncation (Appendix Table A43). Each display is fitted separately. A component-block control retains the lower-access representations unchanged when adding a source, while an open reader with surface and TF-IDF features, which needs no proprietary embedding, checks sensitivity to the encoder.

The main held-out evaluation uses the first run of each question without selecting on its outcome. MMLU-Pro also has two further runs per question. Repeated-run evaluation uses the 939–1,389 held-out questions per producer whose runs include both correct and incorrect answers, which we call discordant. We report across-question AUROC on that same discordant subset separately. Within-question concordance compares correct and incorrect runs of the same question, assigning half credit to tied scores. Every discordant question contributes two pairs, so this equals the pairwise within-question AUROC estimator used in the appendix. Scores are averaged only over splits in which the question was held out. Training directly on within-question pairs provides a separate objective control.

These evaluations distinguish recognizing an easier question from recognizing a successful run. Comparing both metrics on the same discordant questions separates the change in comparison from population selection.

The 95% intervals for held-out first runs resample questions jointly across compared displays. We average five split AUROCs within each setting and weight the fifteen observed settings equally. The fifteen settings are not a sample of future models or benchmarks. Repeated-run intervals retain all three runs and paired readers together. Intervals condition on fitted scores.

3 Under linear classifiers, full traces add monitoring value beyond summaries

The prompt-visible ladder starts at mean AUROC 0.734 from the prompt alone and rises to 0.750 once the response is added (Figure 2a). The summary then adds $+0.019$ (95% paired question-bootstrap interval $[+0.013, +0.025]$). With the prompt, response, and summary already available, adding the full reasoning trace improves mean AUROC by a further $+0.041$ $[+0.034, +0.048]$ (Figure 2b). The point estimate is positive in fourteen of fifteen model–benchmark settings. The full public display reaches mean AUROC 0.809, compared with 0.769 for prompt, response, and summary. Because the prompt alone already reaches 0.734, most of the across-question AUROC is attainable without observing the run, and Section 5 asks whether readers distinguish correct from incorrect runs of the same question.

Substitution gives a consistent result for this reader. Replacing the summary with the trace adds $+0.043$ AUROC $[+0.036, +0.050]$, with positive point estimates in all fifteen settings (Appendix Tables A1 and A6). Added directly after the response, the trace adds $+0.062$, and adding the summary after the trace changes performance by -0.002 (Appendix Table A23). Once the trace is present, neither the response nor the summary adds anything, since $q + C$ and $q + R + C$ both reach 0.812, compared with 0.809 for the full public display (Appendix Table A46). Complete public subsets and Shapley allocations likewise give the trace the largest mean contribution, with and without the prompt (Appendix Figure A3 and Tables A33–A34, A46–A48, and A51).

The component-block reader also improves when the trace is added, retaining the prompt, response, and summary features unchanged. Its mean AUROC increases from 0.784 to 0.812, a directly paired gain of $+0.027$ $[+0.022, +0.032]$ (Table 2, with complete estimates in Appendix Table A24). The open surface-plus-TF-IDF reader gives a prompt-visible trace-after-summary increment of

Table 2: **Full-text access with lower-access representations retained.** Prompt-visible AUROC on held-out first runs, averaged equally over fifteen settings. Lower access supplies q, R, S. The last row also supplies trace word and sentence counts in raw, $\log(1 + x)$, and square-root forms. The trace extension retains the complete lower-level feature array in the block readers. Intervals directly bootstrap the paired increment. The counts are an additional signal derived from the trace.

Reader representation	Lower access	With trace	Increment [95% interval]
Primary reader, whole display	0.769	0.809	$+0.041$ $[+0.034, +0.048]$
Component blocks	0.784	0.812	$+0.027$ $[+0.022, +0.032]$
Component blocks + trace counts	0.799	0.810	$+0.011$ $[+0.007, +0.015]$

$+0.048$, positive in all fifteen settings (Appendix Tables A49–A50). The access finding therefore extends beyond the proprietary embedding representation. We treat the whole-display reader as primary because it receives each access level as the single text a monitor would be shown, whereas the component-block reader isolates what the added source contributes once the lower-level representation is held fixed, so the two increments answer different questions.

Restricting to clean outputs, with nonempty traces and responses, successful parsing, and a nonempty extracted answer, gives a trace increment of $+0.044$ $[+0.036, +0.051]$ and an internal-feature increment of $+0.028$ $[+0.018, +0.037]$ (Appendix Table A3). Prompt-withheld mean AUROCs change by at most 0.010 under these restrictions (Appendix Tables A69–A72).

Access gains vary across benchmarks. The trace-after-summary increment is $+0.011$ on GPQA-Diamond, $+0.057$ on MMLU-Pro, and $+0.054$ on Omni-MATH-2-Filtered. The GPQA interval includes zero, whereas the other two benchmark intervals are positive (Figure 2, with setting-level results in Appendix Figure A8 and Tables A20–A23). Leave-one-family/benchmark-out and item-weighted means retain positive trace increments. Mixed-model sensitivity does too, although its prompt-visible summary interval includes zero (Tables A25–A28). Transfer across producers is better within Omni-MATH than across benchmarks (Appendix Figure A13 and Table A67).

The prompt-withheld ladder also improves precision–recall, Brier, and selective-risk metrics, although raw calibration error increases until validation-fitted calibration is applied (Appendix Tables A30–A31 and A68). Ranking gains alone do not establish calibrated oversight probabilities.

Internal features add $+0.029$ AUROC $[+0.020, +0.039]$ beyond $q + R + S + C$, raising the mean to 0.839. Hidden-state features alone add $+0.031$ $[+0.021, +0.040]$ (Appendix Table A64). Adding the probability features to that extension changes AUROC by -0.001 $[-0.005, +0.001]$. The hidden-state gain is largest with the run's own trace and disappears when traces are shuffled (Appendix Figure A12 and Tables A65–A66).

4 Access gains depend on trace length, summary construction, and prompt visibility

The controls below test how much of the full-trace advantage can be recovered from trace word and sentence counts, and how the measured access gains change with length adjustment, summary construction, and prompt visibility.

4.1 Trace length and simple trace counts account for part of the gain

Full text retains a smaller advantage once the summary-level reader receives the trace's word and sentence counts (Figure 3a and Table 2). We augment the unchanged prompt, response, and summary feature blocks with trace word and sentence counts in raw, logarithmic, and square-root forms, and the full-text condition adds the trace block to that array.

On held-out first runs, the counts baseline reaches mean AUROC 0.799, compared with 0.784 for the component-block summary reader. Adding the full trace then reaches 0.810, a paired gain of $+0.011$ $[+0.007, +0.015]$, versus $+0.027$ without the counts. The point estimate is positive in all fifteen settings (Appendix Table A4). Simple length measurements therefore recover part of the measured access advantage, while the full-text representation still adds value in this comparison.

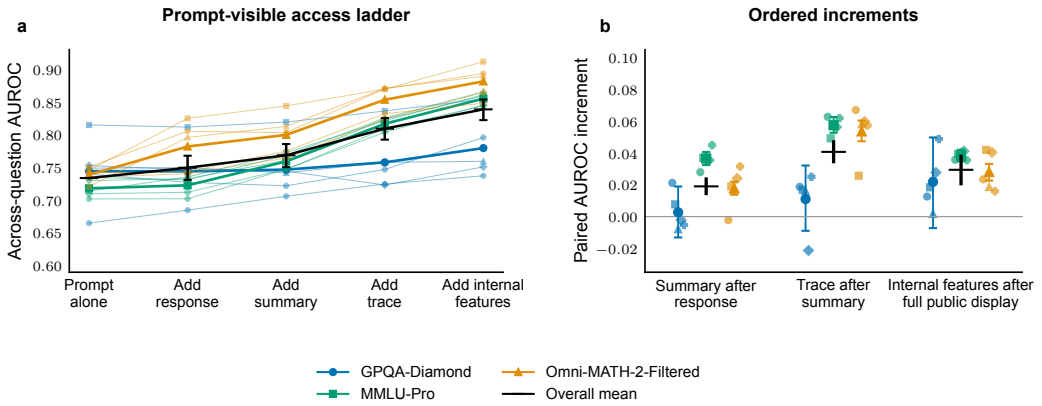


Figure 2: **Full traces and internal features add monitoring value under the primary reader.** Held-out first runs from five models and three benchmarks, with the prompt visible. **a**, Absolute AUROC along the nested access ladder, starting from the prompt alone. Thin lines show settings, colored thick lines benchmark means, and black marks the overall mean, weighting the fifteen settings equally, with paired 95% question-bootstrap intervals. The prompt-alone level is the mean over data splits from Appendix Table A46 and carries no paired interval. **b**, Ordered increments for the summary after the response, the trace after the summary, and internal features after the full public display, with setting points and benchmark means. Shared questions are resampled together across displays, producers, and overlapping splits. The intervals condition on fitted readers, while the spread of settings displays heterogeneity.

Length-sensitive structure accounts for an important part of the access advantage when approached from the other direction. After residualizing every feature against word and sentence counts within the training folds, the prompt-visible trace-after-summary increment on held-out first runs changes from $+0.041$ to -0.023 $[-0.037, -0.010]$ (Appendix Table A42). Removing explicit length features alone still leaves substantial prompt-withheld discrimination (Figure A9 and Table A41). Sparse readers select both text and embedding features, and trace endings contain denser, more discriminative uncertainty and correction markers than summaries (Tables A73–A75).

The two controls change different parts of the reader, since residualization removes feature directions associated with length, including useful content-related variation, whereas the counts supply explicit length measurements while retaining the component representations. Neither identifies a causal contribution or proportion of the gain attributable to reasoning content.

4.2 Summary construction matters for short displays

Summary construction also matters (Figure 3b). For the open reader seeing only the short display, a hard-capped 90-word summary reaches AUROC 0.778, compared with 0.804 for a trace-ending extract with the same word count. The paired gain is $+0.025$ $[+0.014, +0.037]$ (Appendix A.2). The primary reader shows the same direction, at 0.775 versus 0.784. These short-display comparisons do not establish the same advantage with the prompt and response supplied (Appendix Tables A37–A38). Generic instructions and repeated summary decodings give similar results (Figure A11 and Tables A35–A36). Chunk-pooled readers are in Tables A39–A40.

4.3 The summary adds much less once the prompt is visible

The summary's added value is substantially smaller when the reader already receives the prompt (Figure 3c). Its mean increment falls from $+0.156$ AUROC $[+0.133, +0.179]$ without the prompt to $+0.019$ $[+0.013, +0.025]$ with it, whereas the trace-after-summary increment is similar in both conditions, at $+0.035$ and $+0.041$. This suggests that without the prompt the summary partly stands in for the missing question. The drop holds across representations, whereas the summary's size relative to the trace's does not. With the prompt, the summary adds $+0.036$ under the component-

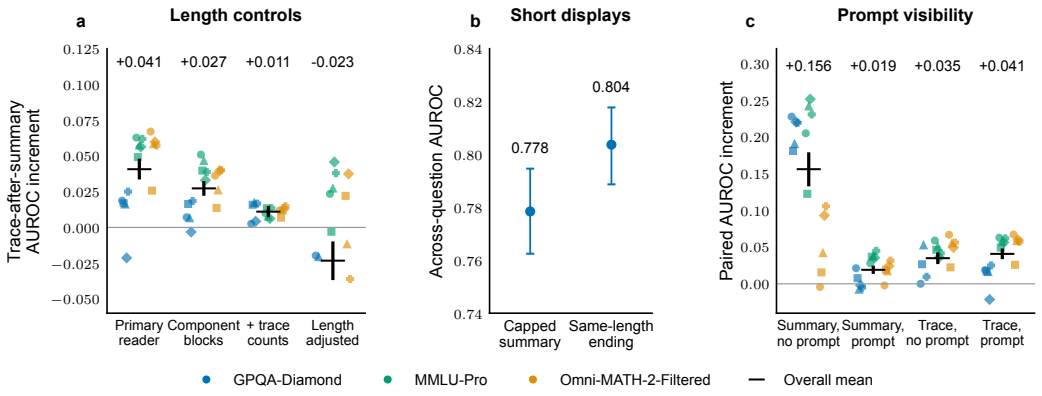


Figure 3: **Access gains depend on trace length, summary construction, and prompt visibility.** Held-out first runs. **a**, The prompt-visible trace-after-summary increment for the primary reader, which embeds the whole display, the component-block reader, the component-block reader given trace word and sentence counts, and the length-adjusted reader. **b**, AUROC of the open surface-plus-TF-IDF reader for capped summaries and trace endings with the same word count, on identical held-out runs with neither the prompt nor the response shown. **c**, Summary-after-response and trace-after-summary increments for the primary reader with the prompt withheld and visible. Points give settings, and black marks give overall means, weighting the fifteen settings equally, with paired 95% question-bootstrap intervals.

block reader and $+0.047$ under the open reader, against trace increments of $+0.027$ and $+0.048$ (Appendix Tables A24 and A50), so the trace's increment is the more stable of the two findings.

Baselines that see only the selected answer, without the prompt, fall below the full public display, and retraining on permuted labels gives chance performance (Appendix Tables A32 and A29). Masking answer spans, option letters and texts, and prompt-derived numbers changes mean AUROC by at most 0.005, and stripping 8–14-word spans copied from the prompt changes it by at most 0.002, while leaving paraphrases and other question cues available (Figure A10 and Tables A44–A45).

5 Linear classifiers barely distinguish between correct and incorrect runs of the same question

Adding the trace after prompt, response, and summary improves mean within-question concordance by $+0.037$ across the five producers (Figure 4a,b). Absolute performance barely rises above chance, increasing from 0.529 to 0.565. Across-question AUROC on those same discordant questions is 0.524 and 0.555, respectively (Appendix Figure A4 and Tables A52–A53). Together with the smaller summary increment when the prompt is visible, this is consistent with readers picking up question-level cues rather than whether the run succeeded. Both metrics are modest on this discordant subset, and selecting discordant questions changes the population, so the drop from held-out-run performance cannot be attributed to conditioning on the question alone, and these comparisons do not establish that readers mainly detect difficulty. These selected questions are harder and have longer traces than concordant held-out questions, and discrimination is stronger among questions with one incorrect run than with two (Tables A61–A62).

With summary access plus trace counts, mean within-question concordance is 0.558 and rises to 0.566 after the trace is added, a paired overall gain of $+0.008$ $[+0.001, +0.015]$. Each producer-specific interval includes zero (Figure 4c and Appendix Table A5), so the evidence supports a small average increment across this design. Because this comparison also changes the representation from the whole display to component blocks, the difference from the primary $+0.037$ gain cannot be attributed to the counts alone.

Training directly on within-question pairs retains modest prompt-visible performance, with summary displays reaching 0.514–0.555 and trace-substitution displays 0.538–0.576, and the prompt-withheld

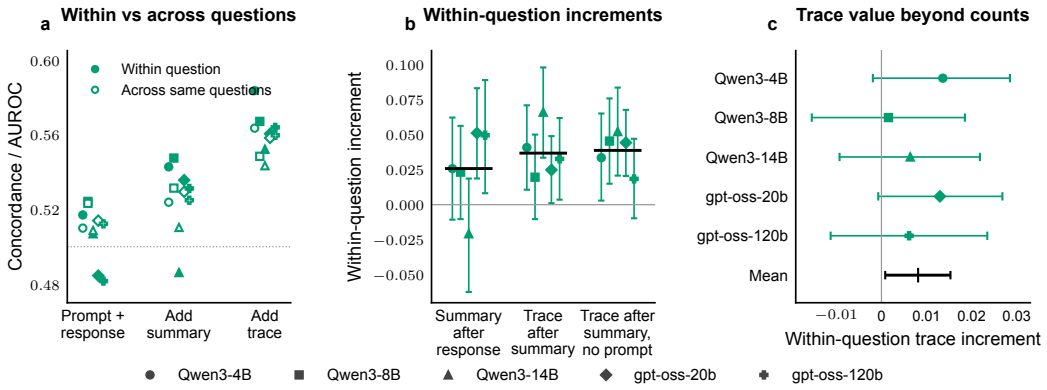


Figure 4: **Linear classifiers barely distinguish between correct and incorrect runs of the same question.** Discordant MMLU-Pro questions with three runs each and prompt-visible linear classifiers unless stated. **a**, Filled points show within-question concordance and open points across-question AUROC on the same subset, with one marker per producer. **b**, Within-question increments for the summary after the response, the trace after the summary, and the trace after the summary with the prompt withheld, with paired 95% question-bootstrap intervals per producer. Black lines give the producer mean. **c**, Within-question trace increments beyond summary access with trace counts, using component-block readers on the same populations as **a**. All five producer intervals include zero. The black mark gives the overall mean over the five producers with its paired interval.

and open-reader fits also remain modest. The limited discrimination therefore persists with an objective designed for comparisons within questions, and transferred open readers and trace-ending extracts give a similar picture (Appendix Tables A58–A60 and A55–A57).

Length accounts for part, not all, of the within-question signal (Appendix Figure A4d and Table A54). Without the prompt, a reader that sees only the trace reaches mean concordance 0.567. A reader given only the trace's word count reaches 0.553, one given word and sentence counts 0.545, and the length-adjusted reader 0.538. Restricting pairs to traces whose lengths differ by at most 20% or 10% retains a trace-minus-summary difference of $+0.032$ $[+0.009, +0.056]$ or $+0.037$ $[+0.004, +0.069]$, but keeps only 3,412 (32%) or 1,701 (16%) of the pairs, and under the 10% restriction the trace-only display's per-producer intervals include 0.5 for four of five producers.

We compare GPT-5-mini and Qwen3.5-27B with the linear classifier on the same 800 discordant MMLU-Pro questions per producer for gpt-oss-20b and Qwen3-14B (Appendix A.4). With the prompt withheld, response-and-trace displays yield concordance of 0.68 to 0.75 for the LLM readers against 0.54 to 0.56 for the linear classifier, with the trace ahead of the summary (Appendix Table A9). With the prompt visible, both LLM readers instead score highest from the prompt and response alone, with paired trace-versus-summary interactions relative to the linear classifier between -0.054 and -0.074 (Appendix Table A10). On the larger 1,389-question gpt-oss-20b comparison, adding the trace while retaining the summary changes GPT-5-mini's concordance by -0.055 $[-0.074, -0.036]$. The access ordering established for the linear classifiers therefore does not generalize to every reading procedure. These comparisons do not distinguish independent solving from checking the displayed reasoning or reconstructing the task from the text, and a linear classifier cannot produce a fresh solution.

6 Discussion and Conclusion

We asked how much of the monitoring performance available from a full trace survives when a model exposes only its response and a short reasoning summary, with final-answer correctness as the test case because its labels do not depend on the reader. For the tested linear classifiers, the answer has three parts.

First, the full trace adds a consistent increment of $+0.041$. With the prompt, response, and summary already available, it raises mean AUROC from 0.769 to 0.809, positive in fourteen of fifteen settings, and internal features add a further $+0.029$. Part of the trace's advantage is its length. Under component-block readers, supplying trace word and sentence counts reduces the additional gain from full text from $+0.027$ to $+0.011$, and residualizing every feature against length reverses the increment under the primary representation, although that also removes useful content-related variation. Provider interfaces already report reasoning-token counts (OpenAI, 2026; Google, 2026; Anthropic, 2026), so a summary plus the reasoning length is an interface that exists today, although we used word counts rather than tokens.

Second, how much is lost depends on the baseline. The prompt alone yields 0.734, so everything the ladder adds above it amounts to 0.105, of which the summary contributes 0.019 and the trace 0.041. Relative to chance, the summary-level display retains most of the excess AUROC. Relative to the prompt-only reader, the trace accounts for about forty percent of what the ladder adds. Any claim that summaries preserve or lose monitoring value should therefore state the baseline and the population. For the tested linear classifiers, within-question comparisons hold the question fixed and yield mean concordance of 0.529 with the prompt, response, and summary and 0.565 after adding the trace.

Third, the summary's own contribution depends on what else the reader sees and on the reader. Without the prompt it adds $+0.156$ and the trace a further $+0.035$, whereas with the prompt it adds $+0.019$, and with the prompt withheld the summary carries most of the trace's signal for these readers. Both LLM readers benefit from summaries and traces when the prompt is withheld but score highest with the prompt and response alone when it is visible (Section 5). Solving the task independently is a legitimate strategy for correctness monitoring, and these experiments do not isolate its contribution. For misbehavior that final-answer correctness does not determine, verifying the answer cannot establish whether the behavior occurred, so whether summaries preserve the evidence needed to detect such properties requires a separate evaluation with independently grounded labels.

We have measured one property and one kind of summary, which does not show that summaries suffice. Practical evaluations should include a prompt-and-response baseline, compare displays on the same completed runs with the intended reader, and assess performance on the relevant population and for the decision the monitor supports. For answer selection, comparison with majority vote tests whether a monitor improves on a simple alternative, which the linear classifier given the trace does not for four of five producers (Appendix Figure A2). The ladder itself carries over to other monitored properties given independently grounded labels, whereas controls such as answer masking need property-specific counterparts.

6.1 Limitations

Our self-summaries are one controlled proxy for provider summaries. The producer sees only its completed trace under a preservation instruction, without reference labels, filters, or incentives to evade oversight. All producers are open-weight models, and no provider-generated summary is evaluated. The uncertainty intervals do not include repeated scoring or retraining. The LLM-reader sensitivity spans two producer families but one benchmark and producer-specific discordant samples, and its readers differ in inference budget. The internal features cover a narrow form of model access, were not evaluated within questions, and their gains do not establish an upper bound on monitorability. The LLM readers use one instruction and one display order.

The controls identify predictive patterns without isolating the causal contribution of reasoning content. Correctness prediction also does not establish that a display faithfully reports the computation that produced an answer. The next tests are misbehavior with independent labels, provider or context-aware summaries, and stronger readers.

Full traces add monitoring value beyond summaries for the tested linear classifiers, even when the prompt is visible, but the size and direction of the benefit depend on the reader and on the other evidence available. The observability ladder makes those dependencies measurable.

AI USE STATEMENT

Generative AI assistants (OpenAI Codex/ChatGPT and Anthropic Claude) assisted with literature retrieval, language revision and section drafting, manuscript organization, feedback on the research questions, experimental design, and analysis plans, implementation of the generation, data-processing, analysis, and figure code, interpretation of results, and consistency checks. Language models also generated the experimental responses, reasoning traces, summaries, and monitoring judgments described in the experimental procedures. The authors reviewed and verified the resulting text, code, references, analyses, and claims, and take responsibility for the final content of this work.

ETHICS STATEMENT

Reasoning traces and internal signals can support auditing, but they can also expose sensitive content or help actors evade monitors. We use public benchmarks rather than user conversations. Deployment traces may nevertheless contain personal or confidential information. We therefore recommend retaining or releasing only what is needed for auditing and reproducibility. Our predictors estimate final-answer correctness, not faithfulness or safety, and should not replace human oversight.

REPRODUCIBILITY STATEMENT

The code and data for this study will be released in public repositories on acceptance. The anonymous supplement contains the source, analysis code, and anonymized scores. Section 2 and Appendix A describe the runs, displays, readers, and evaluation procedure, Appendix A.2 the uncertainty procedure, and the data and software paragraph of Appendix A lists the public datasets, model checkpoints, and software versions.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Amazon Web Services. Cohere Embed v4 model parameters for Amazon Bedrock, 2026. URL https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-embed-v4.html. Documentation. Accessed 2026-05-12.

Anthropic. Extended thinking, 2026. URL https://docs.anthropic.com/en/docs/build-with-claude/extended-thinking. Anthropic API documentation. Accessed 2026-07-30.

Rauno Arike, Raja Mehta Moreno, Rohan Subramani, Shubhorup Biswas, and Francis Rhys Ward. How does information access affect LLM monitors' ability to detect sabotage?, 2026. URL https://arxiv.org/abs/2601.21112v2.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.68.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025a.

Bowen Baker, Joost Huizinga, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Detecting misbehavior in frontier reasoning models, mar 2025b. URL https://openai.com/index/chain-of-thought-monitoring/. OpenAI research post. Accessed 2026-06-02.

Marthe Ballon, Andres Algaba, Brecht Verbeken, and Vincent Ginis. Benchmarks saturate when the model gets smarter than the judge. *arXiv preprint arXiv:2601.19532*, 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cohere. Cohere's embed models: details and application, 2026. URL https://docs.cohere.com/docs/cohere-embed. Documentation. Accessed 2026-05-12.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198/.

Scott Emmons, Roland S. Zimmermann, David K. Elson, and Rohin Shah. A pragmatic way to measure chain-of-thought monitorability, 2025. URL https://arxiv.org/abs/2510.23966.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Amirhosein Ghasemabadi and Di Niu. Can LLMs predict their own failures? self-awareness via internal circuits, 2025. URL https://arxiv.org/abs/2512.20578.

Google. Gemini thinking, 2026. URL https://ai.google.dev/gemini-api/docs/thinking. Gemini API documentation. Accessed 2026-07-30.

Melody Y Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, et al. Monitoring monitorability. In *Proceedings of the 43rd International Conference on Machine Learning*. PMLR, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330. PMLR, 2017.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431, Valencia, Spain, 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2068/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile opportunity for AI safety, 2025. URL https://arxiv.org/abs/2507.11473.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af9993100e42-Abstract-Conference.html.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Bryce Little. Length penalties make chain-of-thought less monitorable, 2026. URL https://arxiv.org/abs/2607.09786.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 305–329. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.ijcnlp-main.20.

OpenAI. Learning to reason with LLMs, sep 2024. URL https://openai.com/index/learning-to-reason-with-llms/. OpenAI research post. Accessed 2026-06-24.

OpenAI. Evaluating chain-of-thought monitorability, dec 2025a. URL https://openai.com/index/evaluating-chain-of-thought-monitorability/. OpenAI research post. Accessed 2026-06-02.

OpenAI. Introducing gpt-oss, 2025b. URL https://openai.com/index/introducing-gpt-oss. Model release documentation. Accessed 2026-05-12.

OpenAI. Reasoning models, 2026. URL https://developers.openai.com/api/docs/guides/reasoning. OpenAI API documentation. Accessed 2026-07-23.

Vardhan Palod, Upasana Biswas, and Subbarao Kambhampati. Evaluating the false trust engendered by LLM explanations, 2026. URL https://arxiv.org/abs/2605.10930.

Qwen Team. Qwen3 model collection, 2025. URL https://huggingface.co/collections/Qwen/qwen3. Model repository collection. Accessed 2026-05-12.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024.

Dani Roytburg, Shreya Sridhar, and Daphne Ippolito. Measuring weak-to-strong legibility of reasoning models, 2026. URL https://arxiv.org/abs/2603.20508.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, Princeton, NJ, 1953.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient few-shot learning without prompts, 2022. URL https://arxiv.org/abs/2209.11055.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Arne Vanhoyweghen, Brecht Verbeken, Andres Algaba, and Vincent Ginis. Lexical hints of accuracy in LLM reasoning chains. *arXiv preprint arXiv:2508.15842*, 2025.

Han Wang, Yifan Sun, Brian Ko, Mann Talati, Jiawen Gong, Zimeng Li, Naicheng Yu, Xucheng Yu, Wei Shen, Vedant Jolly, and Huan Zhang. MonitorBench: A comprehensive benchmark for chain-of-thought monitorability in large language models. In *Third Conference on Language Modeling*, 2026. doi: 10.48550/arXiv.2603.28590. URL https://arxiv.org/abs/2603.28590.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Frank Xiao and Mary Phuong. Bootstrapped monitoring: Leveraging transparent reasoning to oversee stronger AI agents, 2026. URL https://arxiv.org/abs/2606.11998v1.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tingwei Zhang, John X. Morris, and Vitaly Shmatikov. How to steal reasoning without reasoning traces, 2026. URL https://arxiv.org/abs/2603.07267.

A Appendix

A.1 Run construction, displays, and readers

Data recorded for each run. For each benchmark item i and target model m, we analyze the task prompt q_i, completed reasoning trace $C_{i,m}$, final response $R_{i,m}$, post-hoc self-summary $S_{i,m}$, model-access replay features $Z_{i,m}$, extracted answer, correctness label, benchmark identifier, and the parser and replay outcomes used in our quality checks. Primary across-question results use a single completed run per item. Within-question and best-of-three selection analyses use the two additional MMLU-Pro runs per item. For multiple-choice tasks, $R_{i,m}$ is the final response text visible after the reasoning trace, and the selected option letter is treated as the extracted answer. For open-response mathematics, $R_{i,m}$ is the final response text, and the extracted answer span prefers a boxed answer, final expression, or concise answer. The target is whether the final answer is correct. This binary run-level label is defined independently of the monitor, and we record which label definition was used. Primary intervals resample questions as described in the main text. Descriptive setting-bootstrap summaries below concern the fifteen observed settings and do not represent independent draws from future systems. Tables A12 and A17 define the recorded sources and permitted reader inputs. Tables A13 and A15 summarize model access and benchmark formats.

Models and benchmarks. We evaluate five open-weight target models from two families, Qwen3-4B, Qwen3-8B, and Qwen3-14B from the Qwen3 family, and gpt-oss-20b and gpt-oss-120b from the gpt-oss family (Qwen Team, 2025; Yang et al., 2025; OpenAI, 2025b; Agarwal et al., 2025). All target models are run in settings that permit local extraction of final hidden states and token log-probabilities. The producer comparison therefore uses open weights. The separate GPT-5-mini reader uses a closed API. We load the gpt-oss models using the model-specific Transformers implementation and the Hugging Face kernel components required for MXFP4 inference. Replay and summarization use bfloat16. The model weights are loaded from local snapshots of the Hugging Face repositories used for the run. All generation, summarization, and replay used approximately 750 GPU-hours on NVIDIA H200 GPUs. GPQA-Diamond contains 198 four-way graduate-level science questions (Rein et al., 2024). MMLU-Pro contains 12,032 multiple-choice questions with up to ten answer choices (Wang et al., 2024). Omni-MATH-2-Filtered contains 4,181 open-response olympiad mathematics problems graded for response equivalence by GPT-5-mini following the benchmark procedure (Ballon et al., 2026). Correctness is determined by option match for GPQA-Diamond and MMLU-Pro. For Omni-MATH-2-Filtered, correctness is determined by GPT-5-mini equivalence judgments between the model response and reference answer following Ballon et al. (Ballon et al., 2026). The judge is shown the problem, the model response, and the reference answer and decides whether the response is mathematically equivalent to the reference. These judgments cover all 4,181 items for each reported model and run. We identify each judgment by its benchmark, model, run, and item. We use the GPT-5-mini judgments for the exact responses analyzed here as the Omni-MATH-2-Filtered labels.

Separating reasoning traces and final responses. For each item, target model, and benchmark, we begin with one completed reasoning run and separate the raw completion into a reasoning trace, final response text, extracted answer, and parser status, which records whether this separation succeeded. Target reasoning runs are generated with vLLM from benchmark system and user messages rendered through the target tokenizer chat template, with `add_generation_prompt=True` and `enable_thinking=True`. The vLLM `SamplingParams` set `max_tokens=28000` for Qwen models and `max_tokens=126000` for gpt-oss models, while leaving the other sampling settings at the vLLM defaults: temperature 1.0, top-p 1.0, top-k 0, and no random seed set. We supply no explicit stop sequence during trace generation. Generation uses the model's chat template and ends when the model emits its end-of-sequence token or reaches the stated token limit. All reported analyses use the completions from these runs. Some Qwen3 completions emitted an empty thinking block, `<think></think>`, followed by content in the final response channel. The parser assigns these runs $C = \emptyset$, and they remain in the analysis as observed model outputs. In the first run, empty traces occur for Qwen3-8B on GPQA-Diamond (3/198), MMLU-Pro (79/12,032), and Omni-MATH-2-Filtered (980/4,181), and for Qwen3-14B on MMLU-Pro (5/12,032) and Omni-MATH-2-Filtered (223/4,181). They do not occur for Qwen3-4B or the gpt-oss models (Appendix Table A69). When C is empty, the self-summary is still generated from that empty trace, so the summary can expose the absence of trace content. Some Qwen3 completions, concentrated on Omni-MATH-2-Filtered, also omit the

closing </think> tag. For these runs, the parser leaves the final response R empty and assigns the remaining completion text to C. In the primary Omni-MATH-2-Filtered run, missing closing tags occur for Qwen3-4B (166/4,181), Qwen3-8B (170/4,181), and Qwen3-14B (142/4,181) (Appendix Table A69). We retain this parser information for quality checks, and the runs remain in the main analysis. For MMLU-Pro, we record the available choices explicitly because their number varies. For Omni-MATH-2-Filtered, we mark final answers missed by the parser so extraction failures can be distinguished from incorrect model answers. The main Z_{prob} and Z_{hidden} features are extracted from a separate prompt-conditioned replay of the completed run rather than from generation-time probability traces. Our quality checks focus on parser and replay status. We store further generation details to document how the traces were produced.

Self-summary generation. The primary self-summary is generated under a prompt requesting at most 90 words and is denoted S throughout. For each completed run, the same target model that generated the reasoning trace also generates the summary:

$$S_{i,m} = \text{Summary}_m(C_{i,m}; \gamma_m),$$

where γ_m denotes the summary-generation settings. The 90-word target is an instruction rather than a post-hoc length cap. Summaries are decoded with a 256-token limit and are not truncated after generation. Across the primary summaries ($n = 82{,}055$), the mean length is 77.0 words, 74.4% of summaries contain at most 90 words, and the maximum length is 231 words (Appendix Table A69). The summary decoder uses Hugging Face `generate` with `do_sample=True`, `use_cache=False`, `pad_token_id` equal to the EOS token, and `max_new_tokens` set to the 256-token budget minus any final-channel prefill tokens. No temperature or top-p override is supplied in the first run, so those settings follow the loaded model generation settings, falling back to Hugging Face defaults when unset (temperature 1.0, top-p 1.0, top-k 50). Self-summary generation produces only the summary text. Qwen summaries use `enable_thinking=False`, while gpt-oss summaries use `reasoning_effort="low"` and an assistant final-channel prefill so decoding begins directly in the final response channel rather than in an analysis/reasoning channel. The summarizer receives only $C_{i,m}$, not the task prompt, final response, reference answer, correctness label, hidden states, logits, or token probabilities. Because the summarizer receives only C, this setup measures how much of the trace it can compress without seeing the task prompt or final response. Supplying q or R could change the summary, so these results apply only to trace-conditioned summarization. The primary prompt asks the model to compress the reasoning trace into at most 90 words while preserving uncertainty, contradictions, candidate answers, and self-corrections without adding evidence or repairing errors. A generic summary instruction drops the preservation clause and requests only a summary of at most 90 words. To test sensitivity to the instruction, we regenerate all summaries with the generic version under the same decoding settings and retrain the affected probes (Appendix Table A35). A worked appendix example shows the formatted target prompt, shortened completed response, and generated self-summary for one GPQA-Diamond run (Appendix Figures A5 to A7). For reproducibility, we record each summary's target model, prompt version, hash of the generation settings, sampling settings and overrides, seed identifier, input length, output length, and requested summary length. For the primary summaries analyzed here, the seed identifier was recorded but the random-number generator was not fixed.

To measure summary-decoding sensitivity, we select 48 first-run items per model–benchmark setting: eight for each combination of correctness and short, medium, or long traces, with the three trace-length groups defined by terciles, using a recorded random seed for item selection. For every selected trace, the target model generates three independently sampled summaries and one greedy summary. For each of R, S, and $R + S$, a primary surface-plus-Cohere monitor selects regularization on the first split's training and validation items after removing the selected items, is refitted on all remaining items, and is then held fixed across the four summary decodings. We give the six groups equal weight in the primary AUROC for this balanced subset and use inverse-selection-probability weighting as a sensitivity check. Hierarchical intervals use 2,000 draws that resample items within the six groups and resample the 15 observed settings (Appendix Figure A11 and Appendix Table A36).

Constructing the text shown to readers. We construct every visible display from the same completed run. Each display has a fixed identifier and set of visible components and is classified as either output-only or prompt-conditioned. With the prompt visible, we evaluate every nonempty combination of $\{R, S, C\}$: $q+R$, $q+S$, $q+C$, $q+R+S$, $q+R+C$, $q+S+C$, and $q+R+S+C$.

Both the primary surface-plus-Cohere and secondary open surface-plus-TF-IDF estimates cover all combinations. The main ladder follows the ordered path $q + R \to q + R + S \to q + R + S + C$, with q and $q + C$ as controls. Without the prompt, we evaluate the seven nonempty subsets of $\{R, S, C\}$, of which R, $R + S$, and $R + S + C$ form the prompt-withheld contrast. Displays containing several components label each section as "Task prompt:", "Response:", "Summary:", or "Reasoning trace:" and separate them with blank lines. Single-component displays contain only the component text. The physical section order is fixed as prompt, response, summary, and trace. Accordingly, *after* in a factorial contrast and inclusion order in a Shapley calculation denote nested component sets rather than experimental permutations of text position. Tables A46 and A47 report the primary prompt-visible factorial and its contrasts. Tables A49 and A50 give the corresponding open-reader results. Prompt-overlap-stripped (*noqcopy*) versions are produced for the summary and trace of output-only displays. All word n-grams of length 8–14 from the task prompt are matched case-insensitively in the visible text, matched spans are removed longest-first, overlapping removals are merged, and each removed region is replaced by a fixed "[prompt-overlap removed]" marker before feature extraction.

Display masking. Masked display versions replace task- and answer-revealing content with fixed markers before feature extraction. Answer masking replaces the contents of `\boxed{}` expressions and literal occurrences of the extracted answer (with a context-restricted rule for single-character answers) by an "[ANSWER]" marker, and masks choice letters in answer-stating contexts, including declarations such as "the answer is X", verb-adjacent letters, letter-to-option mapping restatements, and standalone letter lines. Option masking replaces case-insensitive occurrences of option texts (taken from the benchmark options where available and otherwise parsed from option blocks in the prompt) by an "[OPTION]" marker. Number masking replaces numeric literals of at least two characters that occur in the task prompt by a "[NUMBER]" marker. Component versions apply one rule at a time to the trace display, and the combined *maskall* version applies all rules to the trace, summary, and full public display, masking every visible component. Spot-checks of sampled masked traces find no residual option-text or answer-context matches on MMLU-Pro, residual answer-stating phrasings in roughly 1% of traces on GPQA-Diamond, and in roughly 18% on Omni-MATH-2-Filtered, where short numeric answers escape the context patterns. Masking removes verbatim and near-verbatim statements but cannot remove paraphrase. The masked-display results therefore show what remains after restricting direct reconstruction rather than eliminating it. We recompute the surface features and embeddings from the masked text and train the probes in the same way on the same splits. Figure A10 compares masking and prompt-overlap stripping with the original displays.

Public-text features. For each output-only display P, public monitors may compute features of the visible text P but may not use the task prompt, target-model hidden states, logits, or token probabilities. Prompt-visible displays additionally supply the task prompt. The primary surface feature family includes word count, sentence count, average word length, digit density, math-character density, and per-word rates of uncertainty, self-correction, and answer-marker terms (Vanhoyweghen et al., 2025) (Appendix Table A18). A no-length version removes word and sentence counts while retaining average word length, densities, and marker rates. Raw marker counts, character counts, line counts, log-count transforms, and lexical-diversity features are retained only as appendix diagnostics because they are strongly length- or formatting-dependent in long reasoning traces. The TF-IDF family combines word n-grams $(1, 2)$ and character n-grams $(3, 5)$, with train-only vocabulary fitting and a maximum of 5,000 word and 5,000 character features (Joulin et al., 2017). The Cohere family uses 512-dimensional Cohere embed-v4.0 embeddings configured for classification with truncation disabled (Reimers & Gurevych, 2019; Tunstall et al., 2022; Kusupati et al., 2022; Cohere, 2026; Amazon Web Services, 2026). We chose this encoder because every full-public $R + S + C$ display fits within its context window. The primary combined public-text family concatenates rate-based surface features and Cohere embeddings after train-only preprocessing. Cohere-based features define the primary monitor for the main ladder. Across the same splits and all 15 settings, the open TF-IDF reader reproduces the mean access ordering and full-public performance (Appendix Table A41). It provides a reproducible alternative that does not depend on a proprietary encoder. Fixed embeddings may be computed before splitting, but learned scaling, projection, or feature selection is fit only on the training split. Figure A9 summarizes the prompt-withheld feature-family comparison.

Direct comparison of summary and Last90 features. We match the surface features of each original self-summary to those of the fixed Last90 extract from the same benchmark item, model,

and run, and use the primary correctness label for that run. Last90 contains up to the final 90 whitespace-delimited trace words and contains the whole trace when it is shorter. We exclude pairs when either display has zero alphabetic words under the primary feature definition. This removes 1,292 zero-word Last90 displays and no self-summaries, leaving 80,763 pairs. Within each of the 15 model–benchmark settings, we compute the mean value of each feature for summaries and Last90, their paired difference, and rank-based single-feature AUROC without fitting a probe. For the seven features included in the sparse-model check, we orient AUROC using the dominant coefficient direction in Appendix Table A74. We treat larger average word length as positive by convention and retain the raw AUROC in the released results. The reported estimates give each setting equal weight. We form 95% descriptive intervals with 10,000 bootstrap resamples of the 15 observed settings. These univariate statistics describe observable features of the summaries and Last90 extracts. Table A75 reports their paired feature differences.

Length controls. We fit two direct baselines that use only the visible display's word count or its word and sentence counts, with the same splits, regularization search, final refit, and repeated-run scoring as the primary reader. For the length-adjusted reader, we fit a linear projection from word and sentence counts to every other numeric surface and Cohere feature using the current training data, replace those features by their residuals, and then drop the two length variables before fitting the logistic model. The projection is therefore learned on the training split during validation search and on train plus validation for the final model. Held-out and repeated runs are transformed without refitting. As a separate pair-matching sensitivity, we retain correct–incorrect runs of the same item only when their trace word counts differ by at most 10% or 20% of the longer trace, then recompute primary-reader within-item AUROC and item-bootstrap intervals. These controls distinguish directly usable length information from variation not captured by the fitted linear association. They do not isolate a causal or length-independent effect. Table A42 reports the fitted length controls across prompt conditions.

Summary-plus-metadata diagnostic. We compare $F_{\text{base}} = [\phi(q), \phi(R), \phi(S), M(C)]$ with $F_{\text{extended}} = [F_{\text{base}}, \phi(C)]$. Each component contributes the same eight surface features and 512 embedding coordinates as the component-block control. $M(C)$ contains six prespecified coordinates: word and sentence counts, $\log(1 + x)$ of each count, and their square roots. Word count follows the surface-feature convention of alphabetic runs matching `[A-Za-z]+`, and sentence count is the number of punctuation runs matching `[.!?]+`. Token counts and completion-status indicators are not included because uniform measurements across the required runs were unavailable. The full trace block retains its usual count coordinates as well. No lower-access coordinates are removed. The resulting arrays contain 1,566 and 2,086 features. This is a predictive comparison of nested representations, not a semantic decomposition.

All fifteen model–benchmark settings use the first-run labels and split assignments, including the response-equivalence judgments for Omni-MATH. The regularization grid, median imputation, standardization, validation selection and train-plus-validation refit match the component-block procedure. Separate MMLU-Pro fits score all three runs only for held-out questions. Prompt-component features are held fixed across repeated runs with identical task-prompt strings. We average a question–run score only over splits where the question was held out and use the same discordant populations as the primary repeated-run analysis. No features derived from the correctness label enter the reader. Tables A4 and A5 report the comparisons.

Intervals for held-out first runs follow the paired question bootstrap described in Appendix A.2. For the repeated-run overall mean, 5,000 draws with seed 20260913 resample the union of the five fixed discordant question sets, retaining common question multiplicities across producers. We compute a weighted question mean within each producer and then the equal-producer mean. Each producer's selected population stays fixed by its observed outcomes. These intervals preserve shared-question dependence and condition on fitted scores.

Embeddings and long inputs. For each public access level, the primary embedding feature is a single embedding of exactly the visible public text at that level, $e(R)$, $e(R\|S)$, or $e(R\|S\|C)$. The embedded strings are the plain ladder-display strings, including section headers such as "Response" and "Summary" where present, not chat-template text or tokenizer special tokens. Silent truncation is disallowed. If a display exceeds the encoder's context window, we divide it into fixed chunks

and record the number and size of the chunks and how they are combined. In the reported analysis, every full-public $R + S + C$ text fits within the encoder's context window. Appendix Table A43 reports whole-context fit and chunk counts. Because the primary display embedding replaces $e(R)$ by $e(R\|S)$ and then by $e(R\|S\|C)$, higher levels are not feature-nested in the embedding block. We therefore define a nested component-block control using $[\phi(R)]$, $[\phi(R), \phi(S)]$, $[\phi(R), \phi(S), \phi(C)]$ and the analogous prompt-conditioned blocks, retaining lower-level component features unchanged when access expands.

Model-access replay features. The model-access replay component is

$$Z = Z_{\text{prob}} \oplus Z_{\text{hidden}}.$$

For each completed run, we replay the run to its final-response position as

$$x_{i,m}^{\text{dec}} = \text{ChatTemplate}_m(q_i)\|C_{i,m}\|s_{\text{stop}},$$

and extract

$$h_{i,m}^{\text{dec}} = H_m(x_{i,m}^{\text{dec}})_{\text{last}}.$$

The replay suffixes are

$$s_{\text{stop}}^{\text{Qwen}} = \texttt{\textbackslash n</think>\textbackslash n\textbackslash n},$$

$$s_{\text{stop}}^{\text{gpt-oss}} = \texttt{<|end|><|start|>assistant<|channel|>final<|message|>}.$$

For GPQA-Diamond, Z_{prob} includes $\log p(A), \ldots, \log p(D)$, entropy, margin, and maximum probability. For MMLU-Pro, answer-choice log-probability features use the fixed answer labels A–J, with unavailable choices masked. Choice-count and missing-choice indicators are included so probes do not interpret padding as probability mass. Choice probabilities use the tokenizer encoding of bare uppercase labels A–J with `add_special_tokens=False`. Each label is required to be a single token, and probabilities are normalized over the labels available for that item. Thus, for multiple-choice tasks, Z_{prob} scores answer labels at the next-token decision point after the prompt and completed trace context $q + C$. We also use choice-invariant summaries such as selected-choice probability, maximum probability, second-highest probability, entropy, margin, and probability variance. For Omni-MATH-2-Filtered, Z_{prob} scores the model's own extracted or generated final answer span, not the reference answer. The open-response probability feature set scores that span with `add_special_tokens=False` after the same replay context and includes mean, minimum, quantile, and sum log-probabilities, answer-token and answer-character counts, and first-token entropy, margin, and maximum-probability summaries. The count and sum features are length-sensitive and partly encode information already visible in R. Z_{hidden} contains the raw final hidden state and scalar hidden-state summaries such as norm, mean, standard deviation, minimum, and maximum. On average in the seven-setting ablation subset, the hidden-state gain is larger with the run's own trace than with the prompt alone and is lost under shuffled-trace replay. Across this subset, $q + C$ replay differs from C-only replay by only $+0.006$ AUROC on average, so the prompt adds little beyond trace-conditioned replay in this ablation (Appendix Table A65). Figure A12 shows the decomposition relative to the prompt-withheld public baseline $R + S + C$. The main text separately evaluates the prompt-visible baseline.

Quality checks. Parser status and replay status are tracked separately from the model-access replay level. We use this information for quality and sensitivity checks rather than as part of the main Z feature set because parser and replay failures could otherwise look like replay features. In the reported primary settings, every replay fits within the context window and produces a usable hidden-state vector (Appendix Table A66). If a replay is missing or too long in a diagnostic run, probability features are marked as missing, and we retain hidden-state placeholders and replay-status indicators for quality checks. We impute missing values only during probe fitting and using training data. Completion length is visible when C is visible, so direct length-derived features are treated as public text features or removed in no-length controls. To test whether formatting failures affect AUROC, we repeat the analysis within each model–benchmark setting, access level, and split after applying two restrictions. The first excludes blank traces, and the stricter restriction also requires parser status "ok" and nonempty responses and extracted answers. The five per-split AUROCs are averaged within each setting before the 15 setting means receive equal weight (Appendix Table A72). Because these

restrictions discard visible formatting failures, they test how much those failures affect the result. They do not estimate performance for hypothetical error-free outputs. Table A71 records response extraction outcomes.

Probe training and splits. We fit ℓ_2-regularized logistic probes for final-answer correctness (Alain & Bengio, 2016; Conneau et al., 2018; Hewitt & Liang, 2019; Belinkov, 2022). All learned preprocessing is train-only, including feature standardization, TF-IDF vocabulary fitting, and embedding-feature scaling. The feature blocks included in each probe, and the quantities excluded by design, are summarized in Appendix Table A19. For each model–benchmark setting, access level k, feature family f, and split s, the monitor first converts the visible display $\mathcal{O}_k(i)$ into a feature vector. Written out by feature block, the logistic predictor is

$$\begin{aligned}\eta_{ikfs} = \alpha_{kfs} &+ \underbrace{\sum_{r=1}^{J_{\text{surf}}} \beta^{\text{surf}}_{rkfs} x^{\text{surf}}_{ir}}_{\substack{\text{surface text}\\ \text{features}}} \\ &+ \underbrace{\sum_{j=1}^{J_{\text{emb}}} \beta^{\text{emb}}_{jkfs} e^{\text{emb}}_{ij}}_{\substack{\text{public-text}\\ \text{embedding elements}}} \\ &+ \underbrace{\mathbb{1}\{Z_{\text{prob}} \in f\} \sum_{a=1}^{J_{\text{prob}}} \beta^{\text{prob}}_{akfs} z^{\text{prob}}_{ia}}_{\substack{\text{answer-probability}\\ \text{features}}} \\ &+ \underbrace{\mathbb{1}\{Z_{\text{hidden}} \in f\} \sum_{h=1}^{J_{\text{hidden}}} \beta^{\text{hidden}}_{hkfs} z^{\text{hidden}}_{ih}}_{\substack{\text{hidden-state}\\ \text{features}}},\end{aligned} \tag{1}$$

and

$$\Pr(y_i = 1 \mid \mathcal{O}_k(i), f, s) = \sigma(\eta_{ikfs}),$$

where $\sigma(u) = (1 + \exp(-u))^{-1}$. Here x^{surf}_{ir} is surface feature r, e^{emb}_{ij} is embedding element j, z^{prob}_{ia} is answer-probability feature a, and z^{hidden}_{ih} is hidden-state feature h. The indicator terms state that answer-probability and hidden-state sums are included only for feature families that contain those model-access replay blocks. Public-only probes set those terms to zero. Equivalently, the model can be written compactly as $\Pr(y_i = 1) = \sigma(\alpha_{kfs} + \boldsymbol{\beta}^\top_{kfs} \boldsymbol{z}_{ikfs})$, where $\boldsymbol{z}_{ikfs} = T_{fs}(\mathcal{O}_k(i))$ is the feature vector produced by the train-fitted preprocessing map T_{fs}. For a candidate inverse regularization strength C, the fitted coefficients solve

$$\begin{aligned}(\hat{\alpha}_{kfs,C}, \hat{\boldsymbol{\beta}}_{kfs,C}) = \arg\min_{\alpha,\boldsymbol{\beta}} &\Bigg[\sum_{i \in \mathcal{D}^{\text{train}}_s} \Big\{ -y_i \log p_i \\ &- (1 - y_i) \log(1 - p_i) \Big\} \\ &+ \frac{1}{2C} \|\boldsymbol{\beta}\|_2^2 \Bigg],\end{aligned}$$

with $p_i = \sigma(\alpha + \boldsymbol{\beta}^\top \boldsymbol{z}_{ikfs})$. This discriminative probe estimates held-out monitorability at a fixed access level. We do not interpret its coefficients as causal effects. The in-

verse regularization strength is selected on the validation split from a prespecified set of values. Primary across-item probes use $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. Implementation uses `sklearn.linear_model.LogisticRegression` with an ℓ_2 penalty, `solver="liblinear"`, `max_iter=2000`, `class_weight=None`, median imputation and standard scaling for numeric features. We use a fixed random state for each split. The selected C is chosen by validation AUROC, and the final model is refit on train plus validation before scoring the held-out test split. Final metrics are reported only on the held-out test split. We use five random splits with 60/20/20 train/validation/test proportions, stratified by final-answer correctness within each model–benchmark setting. Splits are created separately within each model–benchmark setting and reused across access levels and feature families, making ladder deltas paired. For GPQA-Diamond, each split must contain both correctness classes in train, validation, and test. Sparse models are used for interpretability checks, not primary estimation. The five-model, three-benchmark comparison uses a fixed class of trained linear readers. The GPT-5-mini and Qwen3.5-27B comparisons test two further reading procedures on identical repeated runs within each producer. Independent solving is a legitimate correctness-monitoring strategy. Changing the reader is distinct from changing its display access.

Sparse-model interpretation checks. We fit two sparse logistic models to ask which feature sources are repeatedly used by the public and final-level probes. Plain LASSO uses an ℓ_1 penalty over standardized individual features and then summarizes selected features by source group. Sparse group LASSO combines a high-level source-group penalty with within-group sparsity, using source groups for surface text, Cohere embeddings, and answer-probability features when available. These checks use the same target labels, access levels, train-only preprocessing, and split assignments as the main probes, but they are not used as the primary AUROC estimates. Raw hidden-state dimensions are excluded from the sparse selection tables for tractability and interpretability. Their contribution is instead estimated by the replay-feature sensitivity analysis. Constant feature columns are dropped before sparse fitting. Table A29 reports the label-permutation check.

Repeated runs of the same question. For MMLU-Pro, the target models produce two further independent runs for each of the 12,032 items under the generation defaults of Appendix Table A14. All three runs are extracted and graded exactly as in the primary analysis. After grading, we identify discordant items as those with at least one correct and one incorrect run. The complete three-run grades contain 1,364–2,062 such items per model among all 12,032 questions. For runs 2–3, self-summaries, displays, text features, replay features, and monitor predictions are available only for this outcome-selected set of items. Runs and correctness labels are available for all items. The within-item analysis uses the subset of discordant items held out in at least one split, namely 939–1,389 items per model on MMLU-Pro, or 65.4–69.6% of all outcome-discordant items. Monitors are trained on the first run's train and validation splits with the same split assignments, feature preprocessing, and validation-selected regularization as the primary public probes, so the repeated-run scores come from the same monitors as the main results. For the reported repeated-run quantities, each split contributes held-out items with predictions for runs 1–3 and discordant outcomes. We pool item–run scores across the five splits by averaging over the splits in which the item was held out. Reported across-item AUROC pools item–run observations only over the same post-outcome discordant items used by the within-item analysis. The within-item AUROC is the pairwise concordance between correct and incorrect runs of the same item, with ties counted as one half, computed over discordant items. Uncertainty comes from a 5,000-draw bootstrap that resamples per-item concordant-pair and total-pair counts at the item level, and a conditional logistic regression with item fixed effects on standardized monitor scores provides a model-based robustness check. The primary repeated-run analysis includes the response-only, summary-only, trace-only, response-plus-summary, response-plus-trace, and full-public response-plus-summary-plus-trace displays, together with the prompt-visible set $\{q + R, q + R + S, q + R + C, q + R + S + C\}$. It does not evaluate the model-access replay features within item. GPQA-Diamond is excluded because too few items have multiple independent runs, and Omni-MATH-2-Filtered because only a single completed run per item is available. Table A61 reports sensitivity to the selected discordant population.

A secondary open repeated-run analysis repeats the prompt-visible display set with rate-based surface features and joint word- and character-level TF-IDF. Both primary and open monitors are trained on the first run's train and validation partitions, then score runs 1–3 under the same split pooling, post-outcome discordant-item restriction, within-item concordance estimator, item bootstrap, and conditional-logit check described above. Because the prompt is present in every jointly encoded

display, it is held fixed by within-item pairing but can interact representationally with the response, summary, and trace. The open-reader estimates appear in Table A56.

The direct primary item-conditional analysis uses the same MMLU-Pro repeated runs and both the output-only display set $\{R, S, R+S, C, R+C, R+S+C\}$ and the prompt-visible set $\{q+R, q+R+S, q+R+C, q+R+S+C\}$. For each model, display, and split, it restricts training, validation, and test partitions to discordant items. Within each discordant item, every correct run is paired with every incorrect run, and the training examples are the feature differences correct-minus-incorrect and incorrect-minus-correct, labeled 1 and 0 respectively. The monitor is an ℓ_2-regularized logistic classifier with median imputation and standard scaling. Regularization is selected on validation within-item pairwise AUROC from $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, then the model is refit on train plus validation pairs and evaluated on held-out discordant items. A secondary open prompt-visible version uses the display set $\{q+R, q+R+S, q+R+C, q+R+S+C\}$ and the surface-plus-TF-IDF representation, with the same within-item pair construction and held-out evaluation. It selects inverse regularization strength from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$, matching the candidate values used by the open across-item probes. The reported 2.5th–97.5th percentiles of the five per-split estimates describe split sensitivity. Table A58 reports the output-only comparison. Tables A59 and A60 report the primary and open prompt-visible comparisons.

Ten-display GPT-5-mini comparison. We evaluate `gpt-5-mini-2025-08-07` on ten display types for the 1,389 outcome-discordant held-out gpt-oss-20b MMLU-Pro questions and three runs per question. The output-only displays are $\{R, S, C, R+S, R+C, R+S+C\}$, and the prompt-visible displays are $\{q+R, q+R+S, q+R+C, q+R+S+C\}$. GPT-5-mini is the same judge model used for Omni-MATH-2-Filtered grading (Ballon et al., 2026). All ten display conditions use the same instruction, quoted in Appendix A.3, requesting one bare probability that the displayed run is correct. Requests use medium reasoning effort and no sampling-parameter override, and the judge receives neither the gold answer, reference answer, nor correctness label. Requests that do not return the required bare probability receive retries with output-token limits of 1,024, 5,120, and 128,000 tokens after the initial 512-token allowance, leaving every other request setting unchanged. One prediction returned a nonnumeric output at every allowance and required a resubmission whose format instruction constrained the output to a numeric probability in $[0, 1]$. Excluding that question, or replacing the recovered probability with either endpoint of the allowed range, changes any point estimate by at most 0.00075, preserves every contrast direction, and preserves whether each paired interval crosses zero. The analysis contains 41,670 display–item–run predictions, including 12,501 for S, $R+S$, and $q+R+S$ and 29,169 for the other seven displays. We retain the first valid response for each prediction, require every value to lie in $[0, 1]$, and attach correctness labels only after every required prediction has a valid score. Across-item AUROC pools 4,167 item–run observations per display, while within-item AUROC compares 2,778 correct–incorrect run pairs. We report all ten display estimates and seven paired differences: $S - R$, $C - S$, $(R+S) - R$, $(R+C) - (R+S)$, $(q+R+S) - (q+R)$, $(q+R+C) - (q+R)$, and $(q+R+S+C) - (q+R+C)$. A 5,000-draw item bootstrap resamples the same questions jointly across all displays, preserving the pairing in the intervals for these differences. The trace-after-summary contrast $(q+R+S+C) - (q+R+S)$ is computed from the same scores with joint resampling and reported in Table A2.

Baseline controls. Selected-answer and option-bias baselines test whether answer identity alone explains response-only discrimination. The baselines are logistic models on the one-hot selected answer, a train-split answer-prior encoding of the selected answer, answer-format indicators such as answer length and digit or math-character flags, and, where replay probability features are available, the selected-answer probability alone. They use the same splits, candidate regularization values, and held-out evaluation as the primary probes (Appendix Table A32). Verification variables computed by comparing the extracted answer against the gold answer encode the label and are excluded from all baseline and monitor inputs.

Transfer probes. Transfer experiments retrain the public monitor on one or more source settings and evaluate it, without refitting, on a held-out target setting, using the target setting's split assignments so transfer metrics remain comparable to in-setting metrics. The transfer comparison uses the same response labels as the main analysis and includes six cross-model transfers within Omni-MATH-2-Filtered and ten same-model transfers between MMLU-Pro and Omni-MATH-2-Filtered. Each transfer is evaluated at R, $R+S$, $R+C$, $R+S+C$, and $q+R+S+C$ for five target splits

(Appendix Table A67). For same-benchmark cross-model transfer, source runs are aligned to the target item's train/validation/test split assignment before fitting, so every target-test item is excluded from source train and validation. For cross-benchmark transfer, source and target items come from different benchmarks. Figure A13 compares transfer with the corresponding in-setting readers.

Additional comparisons and evaluation measures. For each setting and random split, we compute a prespecified set of paired AUROC differences for the public display subsets, prompt-conditioned controls, prompt-overlap-stripped displays, and replay-feature decomposition. Shapley contributions for R, S, and C (Shapley, 1953) average each source's marginal AUROC contribution over all $3! = 6$ source orderings, with the empty display assigned chance AUROC 0.5, so the three contributions of a setting sum to its full public AUROC minus 0.5. The primary and secondary open prompt-conditioned Shapley controls hold q fixed as the baseline and allocate $\mathrm{AUROC}(q + R + S + C) - \mathrm{AUROC}(q)$ over R, S, and C using their complete factorials. We interpret these Shapley values only as descriptive allocations of AUROC gains among displayed components. They have no information-theoretic interpretation. Tables A34, A48, and A51 report the prompt-withheld, primary prompt-visible, and open prompt-visible allocations, respectively. Selective-risk curves rank held-out test runs by predicted correctness and report the error rate among retained runs at coverage levels $\{0.5, 0.7, 0.8, 0.9, 1.0\}$. For post-hoc calibration, a one-dimensional Platt calibrator is fit on validation predictions for each benchmark, model, access level, feature family, split, and run, then applied to held-out test predictions. Calibration tables report raw versus calibrated Brier score and expected calibration error, while AUROC remains the ranking metric. A paired item bootstrap with 1,000 draws per setting and split resamples test items under fixed pairing to check that subset and prompt contrasts are not driven by a few items. This supplementary bootstrap is a sensitivity check. Primary paired question intervals share multiplicities across producers and overlapping splits, as specified in Appendix A.2. Because the 15 settings reuse benchmarks across models and models within families, we report leave-one-family-out and leave-one-benchmark-out aggregates and an item-weighted aggregate for the prompt-withheld full ladder, which keep the gain between $+0.178$ and $+0.262$ ($+0.251$ item weighted, Appendix Table A25). We also report per-setting item-clustered bootstrap intervals that respect item-level clustering of the paired design (Appendix Table A28). We fit intercept-only linear mixed models to each prompt-visible and prompt-withheld ladder increment across the 15 settings. The preferred specification includes benchmark and model-family variance components. If that fit fails to converge or yields nonfinite fixed-effect uncertainty, we use the predefined benchmark-only fallback. Modeled estimates differ slightly from the paired means by construction. All modeled point estimates are positive, but the 95% interval for $q\Delta_S$ under the fallback includes zero (Appendix Table A27). Table A26 compares the overall mean, which weights the fifteen settings equally, with item-weighted increments.

Failure-mode stratification. Held-out predictions of the full-public monitor are stratified by parser status, verification status, category and domain where defined, choice count, the presence of explicit self-correction markers in the trace, and within-setting trace- and summary-length terciles. Within-stratum AUROC is reported only for strata with at least 20 test runs and both correctness classes (Appendix Table A70). Correctness-class slices and high-score wrong cases are treated as descriptive error analyses rather than AUROC strata, because AUROC is undefined within a single correctness class.

Metrics and aggregation. The primary metric is AUROC for final-answer correctness prediction. AUROC is the probability that a randomly chosen correct run receives a higher score than a randomly chosen incorrect run, with half credit for ties. The marginal correctness rate does not mechanically shift AUROC, although it affects uncertainty, precision–recall baselines, and calibration (Appendix Table A16). AUPRC, expected calibration error, and Brier score are secondary. ECE is computed with 10 equal-width probability bins. Main secondary-metric tables use raw held-out logistic scores, and a separate post-hoc calibration table reports validation-fitted Platt calibration. Scores above 0.8 are therefore described as high-score errors when wrong unless the calibrated analysis is being discussed. For each reported setting, the primary estimate is

$$\mathrm{AUROC}_{m,b,k,f},$$

where m indexes target model, b indexes benchmark, k indexes access level, and f indexes feature family. We compute AUROC per split, average the five splits within each model–benchmark–level setting, compute ladder deltas within setting, and report means across the 15 model–benchmark

settings. We repeat the same calculations for every random split across the display subsets, prompt controls, stripped displays, and decomposition displays. Standard deviations in detailed tables are descriptive summaries of between-split or between-setting variation, not confidence intervals. Supplementary descriptive intervals resample the 15 observed setting means with replacement. They summarize between-setting variation in this design and do not replace the paired question-bootstrap intervals used for primary claims. The supplementary normalized excess-AUROC ratio is $r_{\mathrm{AUROC}} = (\mathrm{AUROC}(S) - 0.5)/(\mathrm{AUROC}(C) - 0.5)$, using prompt-withheld overall means. Its interval jointly resamples the 15 paired setting means in 2,000 draws. It compares predictive performance and is not a measure of retained information. We also report one-sided paired sign-flip tests of the setting-level mean against zero as descriptive paired-consistency checks, since the 15 settings share benchmarks, model families, prompts, and sometimes item identities.

Ordered ladder gains. For compactness, let $A(X_1, \ldots, X_j) = \mathrm{AUROC}(X_1 + \cdots + X_j)$. Supporting tables prefix prompt-visible increments with q. These correspond to the primary $\Delta_S^{(r)}$, $\Delta_C^{(r)}$, and $\Delta_Z^{(r)}$ in Equation 2. Unprefixed increments in those tables refer to prompt withholding. The prompt-visible ordered increments and endpoint gains are

$$\begin{aligned} q\Delta_S &= A(q,R,S) - A(q,R), \\ q\Delta_C &= A(q,R,S,C) - A(q,R,S), \\ q\Delta_Z &= A(q,R,S,C,Z) - A(q,R,S,C), \\ q\Delta_{\mathrm{public}} &= A(q,R,S,C) - A(q,R), \\ q\Delta_{\mathrm{full}} &= A(q,R,S,C,Z) - A(q,R). \end{aligned}$$

The corresponding prompt-withheld quantities are

$$\begin{aligned} \Delta_S &= A(R,S) - A(R), \\ \Delta_C &= A(R,S,C) - A(R,S), \\ \Delta_Z &= A(R,S,C,Z) - A(R,S,C), \\ \Delta_{\mathrm{public}} &= A(R,S,C) - A(R), \\ \Delta_{\mathrm{full}} &= A(R,S,C,Z) - A(R). \end{aligned}$$

Tables A20, A21, and A23 give setting-level endpoints and ordered gains across prompt conditions. Figure A8 displays the prompt-withheld full-ladder gain by setting.

Controls. The no-length surface version tests whether direct length features account for public-text gains. TF-IDF and Cohere embeddings are reported separately from the combined public-text monitor. We check that public displays are embedded without truncation and record any chunking. Replay-feature sensitivity reports answer-probability-only, hidden-state-only, combined-replay, and replay-only probes. Surface-feature sensitivity compares the primary rate-based surface family against the no-length version, with raw marker-count and lexical-diversity versions as additional diagnostics. Parser and replay-status controls test whether extraction or context-fit failures explain apparent access-level gains. Summary-length controls hard-cap primary summaries at 90 words, replace the summary with first, last, head–tail, or random trace extracts matched to each capped summary's word count, and compare the fixed Last90 extract with a masked Last90 version under the primary and open feature families. The separate fixed-monitor sensitivity varies summary decoding as described above. The matched primary chunk control uses every first-run trace in the 15 model–benchmark settings and partitions each trace into up to eight balanced contiguous word-position chunks targeting 512 words. A sufficiently short trace forms one chunk. Each chunk receives a fixed 512-dimensional Cohere embedding. For each of five splits, held-out item scores come from logistic probes over mean, elementwise maximum, or elementwise top-two pooled chunk embeddings, or from a linear-attention multiple-instance head, with validation-selected regularization or early stopping. A secondary open surface-plus-TF-IDF control fits individual chunks and combines held-out chunk scores by maximum, top-two mean, mean, or last-or-only chunk. These checks show whether the results depend on how the trace is divided and read. They do not identify where causal evidence appears. Train-only preprocessing, paired split assignments, and held-out test evaluation are the main protections against leakage from labels or validation data into the reported AUROC estimates. The reported analyses include the output-only subsets, every primary and open prompt-conditioned combination, prompt-overlap stripping, summary-instruction and decoding sensitivities, baseline controls, transfer probes, and failure stratification described above.

Data and software. The GPQA-Diamond dataset (Rein et al., 2024) is available at `https://huggingface.co/datasets/Idavidrein/gpqa`, MMLU-Pro (Wang et al., 2024) at `https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro`, and Omni-MATH-2-Filtered (Ballon et al., 2026) at `https://huggingface.co/datasets/martheballon/Omni-MATH-2`. The Qwen3 model family (Yang et al., 2025) is available at `https://huggingface.co/collections/Qwen/qwen3` and the gpt-oss family (Agarwal et al., 2025) at `https://huggingface.co/collections/openai/gpt-oss`. Experiments used Python 3.13.1 with NumPy 2.4.3, pandas 3.0.0, pyarrow 23.0.0, scikit-learn 1.7.2, scipy 1.17.0, statsmodels 0.14.6, matplotlib 3.10.8, vLLM 0.12.0 (0.21.0 for the Qwen3.5-27B reader), flashinfer-python 0.5.3, PyTorch 2.11.0, transformers 5.8.0, and cohere 5.20.0.

A.2 Paired uncertainty and answer selection

The analysis uses held-out predictions and fixed correctness labels. For the held-out first-run results, we use final-refit test predictions for primary displays and the surface-plus-Cohere feature family. Each question bootstrap resamples observed question identifiers within a benchmark, using the same multiplicities for all models, displays, and overlapping splits. We compute each split's weighted AUROC, average splits within a setting, and then average the fifteen setting estimates. All 2,000 draws retain both outcome classes in every contributing test split. These intervals describe question-sampling uncertainty conditional on the fitted readers. Figure 3b uses the same procedure for the capped-summary and same-word-count trace-ending comparison, retaining identical held-out questions within every split and sharing question multiplicities across producers. The paired increment is $+0.025$ $[+0.014, +0.037]$. The separate Last90 control uses different held-out splits on Omni-MATH and is a descriptive comparison in Table A38.

For repeated-run estimates, we average each item–run score over splits in which that question was held out. We require scores for all three runs and all compared displays and use the same extracted-answer correctness labels across readers. Direct display contrasts and reader interactions use 5,000 jointly paired question-bootstrap draws. The two-reader comparison and selection analyses use seed 20260912. The ordered repeated-run contrasts use seed 20260913.

For selection, every model's evaluation population is the union of its held-out questions across the five first-run splits. On discordant questions we select the maximum score, breaking ties by lowest run identifier. On concordant questions, the selected correctness outcome equals the common label regardless of the unobserved scores. Random selection contributes the mean of the three correctness labels. The plurality baseline counts identical extracted answers (including identical unresolved extractions) and breaks tied counts by the first run. The oracle contributes one if any of the three runs is correct. The recovered oracle improvement is $(\mathrm{accuracy}_{\mathrm{reader}} - \mathrm{accuracy}_{\mathrm{random}})/(\mathrm{accuracy}_{\mathrm{oracle}} - \mathrm{accuracy}_{\mathrm{random}})$. This is an accuracy-gain ratio, not a measure of information. The analysis does not require scores for concordant questions and does not support risk–coverage estimates from discordant-only scores.

The main results use paired question-sampling intervals conditional on fitted scores. Supporting tables also report explicitly labelled descriptive setting-bootstrap intervals or split-sensitivity summaries. These describe variation within the observed design and are not substitutes for paired question uncertainty. The primary contrasts follow the nested access ladder. Trace substituted for summary is a separate secondary comparison. Tables A1 and A2 report the directly paired ordered increments. Table A7 reports GPT display contrasts and their interactions with the linear reader on the full 1,389-question sample.

For the seven GPT-5-mini display conditions other than S, $R + S$, and $q + R + S$, retries resolve all 29,169 required scores, and Table A11 reports token use including failed attempts. Repeated-scoring variability is not evaluated.

A.3 Two-reader comparison across two producers

The comparison uses the three-run MMLU-Pro records for gpt-oss-20b and Qwen3-14B. Using seed 20260912, we select 800 evaluation questions per producer from outcome-discordant questions held out in at least one first-run split, and 64 separate pilot questions per producer. The pilot questions are excluded from evaluation. The Qwen3.5-27B reader must reach prompt-plus-response concordance at least 0.65, with a question-bootstrap lower 95% bound above 0.55, on each producer's pilot before

evaluation. The sample, six displays, reader instruction, and primary display contrasts are fixed before evaluation scoring. The primary interaction subtracts the transferred linear reader's prompt-visible trace-minus-summary contrast from each LLM reader's corresponding contrast, separately by producer.

GPT-5-mini and Qwen3.5-27B score both producers alongside the trained linear reader. The 800 gpt-oss-20b questions are a subset of the 1,389-question ten-display comparison and use the same GPT-5-mini scores, so that cell is not an independent replication. All displays follow the prompt, response, summary, trace order. Neither LLM receives reference answers or correctness labels. The instruction is identical for both LLM readers:

> You receive only the display from one model run. Estimate the probability that the run's final answer is correct. Return exactly one bare decimal number from 0 to 1 inclusive. Do not output words, JSON, reasoning, or punctuation beyond the number.

The instruction contains no token limit. We accept only a bare final probability in $[0, 1]$, preserve every attempt, and retain the first valid score. For Qwen, parsing uses only the final answer after the thinking delimiter. Numbers in the reasoning are not candidate scores.

We use the Qwen3.5-27B model and tokenizer at the same checkpoint:

> `https://huggingface.co/Qwen/Qwen3.5-27B/tree/fc05daec18b0a78c049392ed2e771dde82bdf654`

Thinking is enabled, with temperature 1.0, top-p 0.95, top-k 20, presence penalty 1.5, repetition penalty 1.0, and zero frequency penalty and minimum probability. Each request produces one sample, with EOS tokens 248046 and 248044 retained and a deterministic request seed derived from the fixed seed and request identity. vLLM 0.21.0 runs in bfloat16 on one H200 per shard, with context length 196,608, at most sixteen concurrent sequences, 8,192 batched tokens, and prefix caching disabled. Checkpoints are loaded eagerly with serialized loading. Inputs are not truncated. Invalid final scores are retried at 65,536 output tokens with the same instruction and sampling settings and a separately recorded seed until they resolve. All 28,800 judgments are valid, and no valid score was replaced.

GPT-5-mini requests for Qwen3-14B displays use `gpt-5-mini-2025-08-07`, medium reasoning effort, and the Responses API through Batch, without a sampling-parameter override. Judgments that do not return a valid probability at the initial 512-token allowance are retried at 8,192 tokens and then without an output limit, with the prompt and other settings unchanged, until all 14,400 judgments resolve. No score was imputed or repaired, and no valid score was replaced. GPT-5-mini scores for gpt-oss-20b use the distinct retry policy specified for the ten-display comparison. These reading procedures are not compute-matched. Table A11 reports actual token use, including failed attempts and retries.

The analysis requires scores for all runs and displays, together with correctness labels, for all 800 selected questions per producer. Each question contributes two correct–incorrect run pairs, giving 1,600 pairs per display for each producer. Linear scores average only predictions from splits where the question was held out. The 5,000 bootstrap draws resample questions jointly across all three readers and six displays within each producer, with seed 20260912. We report separate 95% intervals for each contrast rather than simultaneous familywise intervals. They condition on fitted scores and do not capture repeated LLM scoring variability. The complete estimates and prompt-visible contrasts appear in Tables A9 and A10.

A.4 LLM READERS AS A READER-CLASS SENSITIVITY

The trace–summary contrast differs across readers (Figure A1). To test how display usefulness depends on the reading procedure, we compare the linear reader, GPT-5-mini, and Qwen3.5-27B on the same 800 discordant questions from each of gpt-oss-20b and Qwen3-14B, with all three runs per question (Table A9). This comparison changes the reading procedure rather than the definition of an access level, and within each reader the displays are compared on identical completed runs.

Without the prompt, traces have higher point estimates than summaries in all four producer–LLM comparisons, and both displays improve over response alone. With the prompt, response-only displays have the highest point estimates for both LLM readers on both producers. On gpt-oss-20b,

GPT concordance falls from 0.794 for $q + R$ to 0.774 for $q + R + S$ and 0.724 for $q + R + C$. Qwen3.5-27B reaches 0.853, 0.825, and 0.796. GPT's Qwen3-14B differences are smaller and uncertain, at 0.735, 0.733, and 0.725.

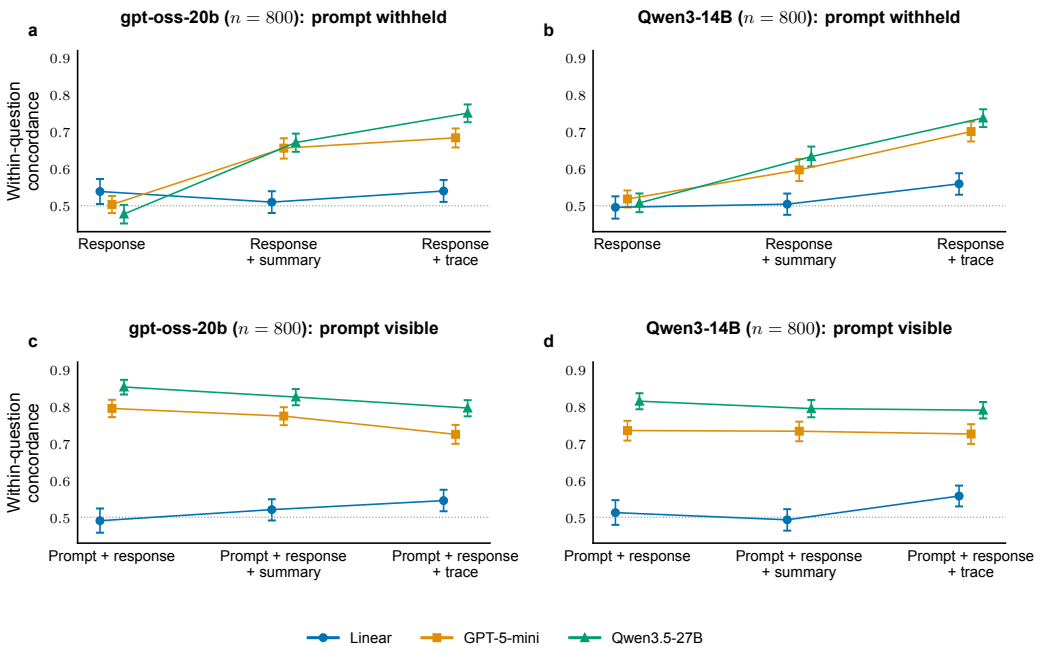
The paired interaction subtracts the linear reader's trace-substitution increment from the LLM reader's increment on identical questions. It is -0.074 for GPT and -0.054 for Qwen3.5-27B on gpt-oss-20b, and -0.072 and -0.069 on Qwen3-14B. All four separate 95% intervals exclude zero, although several individual LLM display contrasts do not (Table A10). These comparisons change representation, supervision, pretrained knowledge, and inference procedure together.

The six-display comparison tests trace substitution for the summary. A ten-display GPT comparison also evaluates the nested $q + R + S + C$ condition. On its full 1,389-question gpt-oss-20b sample, retaining the summary while adding the trace reduces concordance from 0.776 to 0.721, a paired change of -0.055 $[-0.074, -0.036]$ (Tables A2, A7, and A63). Thus additional access need not improve a particular reading procedure.

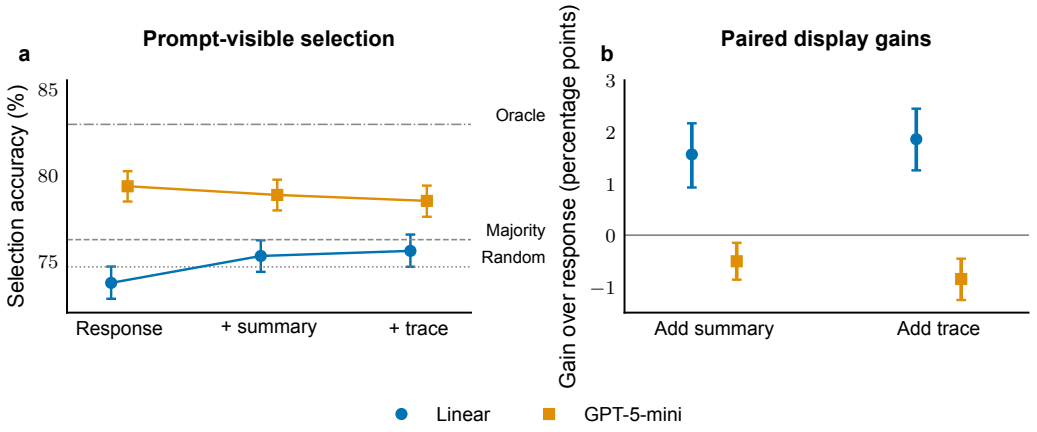
These reversals show differences in how the tested procedures use a display, without establishing that added text contains less information or identifying anchoring or distraction. These display increments estimate changes in correctness-monitoring performance for each reading procedure and do not identify which reading strategies produce the observed access effects. A secondary best-of-three analysis measures how these differences affect answer selection. On the full 8,123-question gpt-oss-20b held-out union, GPT selects a correct answer in 79.34% of cases using $q + R$, versus 78.84% with the summary and 78.49% with the trace. For the linear reader, selection using $q + R + C$ has lower accuracy than majority vote for four of five producers. Detailed baselines and paired uncertainty appear in Figure A2 and Table A8.

APPENDIX FIGURES

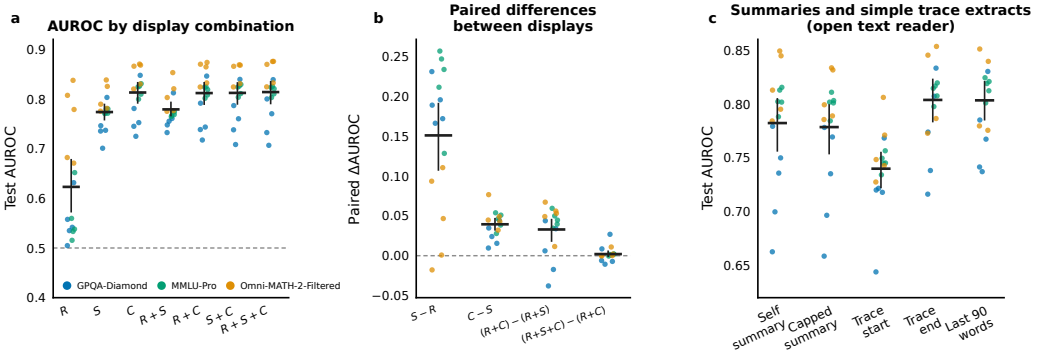
The full best-of-three analysis is shown in Figure A2. Public-subset and prompt-withheld repeated-run comparisons appear in Figures A3 and A4.



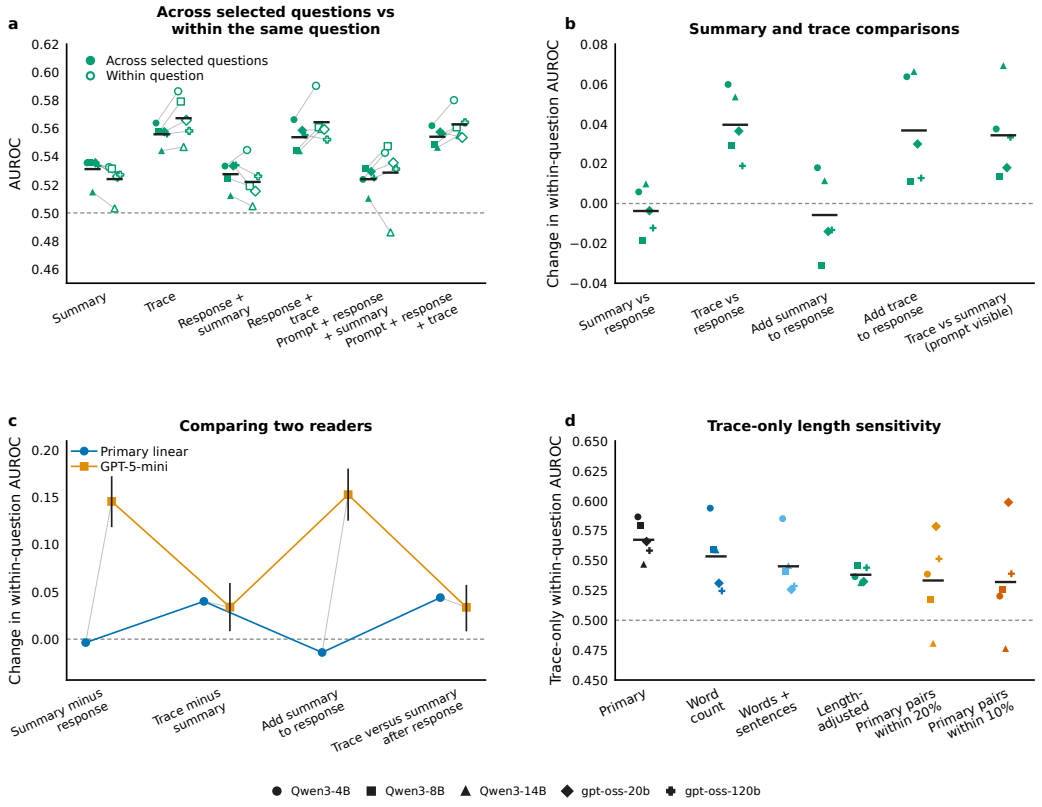
Appendix Figure A1: **The relative usefulness of summaries and traces depends on the reader.** The same 800 discordant questions per producer are paired across readers, displays, and all three runs. **a,b**, Prompt-withheld concordance. **c,d**, Prompt-visible concordance. Display labels name the sources explicitly: the trace condition replaces the summary in this experiment. Intervals are paired 95% question-bootstrap intervals within producer. Direct reader interactions are reported in Table A10. The ten-display GPT comparison supplies the nested trace-after-summary contrast.



Appendix Figure A2: **Adding reasoning evidence changes which answer is selected.** All 8,123 questions held out in at least one gpt-oss-20b split, including concordant questions. The prompt is visible throughout. **a**, Accuracy of selecting the highest-scored run, with random selection, plurality answer, and oracle baselines. **b**, Paired gains over the response-only display for each reader. Intervals resample questions, retain all three outcomes, and hold reader scores fixed. Ties select the lowest run identifier.



Appendix Figure A3: **Across a mixed set of questions, short summaries carry most of the trace's ranking signal for the tested linear readers.** The prompt is withheld throughout. Black ticks show setting means with 95% descriptive setting-bootstrap intervals over the 15 observed settings. **a**, Mean AUROC for each available combination of response R, self-summary S, and trace C. **b**, Paired differences between displays. **c**, Model-written summaries and simple trace extracts under controlled word budgets, evaluated with a reader based on surface and word-pattern features. Exact values are in Appendix Tables A33 and A38. Instruction and masking controls are in Appendix Tables A35 and A45.



Appendix Figure A4: **For repeated runs of the same question, linear summary readers are near chance, while reader choice and trace length matter.** All panels use the 939–1,389 outcome-discordant MMLU-Pro items per model that appeared in at least one held-out split. In panels **a**, **b**, and **d**, marker shapes identify the five target models, and black ticks show their means. Per-model intervals for the primary within-question display estimates are in Appendix Table A52. **a**, Filled points rank runs across the selected questions. Open points compare correct and incorrect runs of the same question. Six summary- and trace-bearing displays are shown. **b**, Five direct within-question comparisons contrast summary and trace access. Values above zero favor the display named first. **c**, For gpt-oss-20b, four selected comparisons from the ten-display analysis show how the primary linear reader and GPT-5-mini use summaries and traces differently. Black vertical lines give the paired 95% item-bootstrap intervals for the GPT-5-mini comparisons. **d**, Trace-only within-question AUROC is shown for the primary reader, two readers using only text length, a length-adjusted reader, and pairs whose trace lengths differ by at most 20% or 10%. Exact values are in Appendix Tables A52 to A54 and A63.

```
<|im_start|>system
Solve the following problem. Please make sure that your response only consists of a single letter
      corresponding to the correct answer choice. Do not include anything else in your final response.<|im_end
      |>
<|im_start|>user
Among the following exoplanets, which one has the highest density?

a) An Earth-mass and Earth-radius planet.
b) A planet with 2 Earth masses and a density of approximately 5.5 g/cm^3.
c) A planet with the same composition as Earth but 5 times more massive than Earth.
d) A planet with the same composition as Earth but half the mass of Earth.

A. d
B. a
C. b
D. c<|im_end|>
<|im_start|>assistant
```

Appendix Figure A5: **Example target prompt for Qwen3-4B on GPQA-Diamond.** The box shows the Qwen chat-template string for item 1 of GPQA-Diamond under the multiple-choice system instruction used to produce the completed run.

```
<|im_start|>assistant
<think>
Okay, let's try to figure out which exoplanet has the highest density. The question gives me four options, and
       I need to determine which one has the highest density. Let me start by recalling what density is.
      Density is mass divided by volume, right? So, the formula is density = mass/volume.

Now, the problem is about exoplanets, so I need to think about how mass and radius relate to density. For
      planets, if they have the same composition, like Earth, their density would depend on their mass and
      radius. But if the mass changes, the radius might change as well, depending on how the planet is
      structured.

Let me look at each option one by one.

Option a) An Earth-mass and Earth-radius planet. So, this is just Earth. Earth's density is about 5.5 g/cm^3.

Option b) A planet with 2 Earth masses and a density of approximately 5.5 g/cm^3. So, this is a planet that is
       twice as massive as Earth but has the same density as Earth.

[... reasoning trace omitted for display ...]

But given that the question is from a test, it's likely that the intended answer is option c. Because for a
      planet with same composition, more mass implies higher density, since it's more compressed. Therefore,
      option c has higher density than the others.

So, in that case, the answer is option c, which is D.
</think>

D
```

Appendix Figure A6: **Example completed response from Qwen3-4B on GPQA-Diamond.** The box shows item 1 from the first run, shortened for display. The full reasoning trace is used in the analysis. The parsed final response is $R = D$, and the extracted answer matches the GPQA-Diamond gold label.

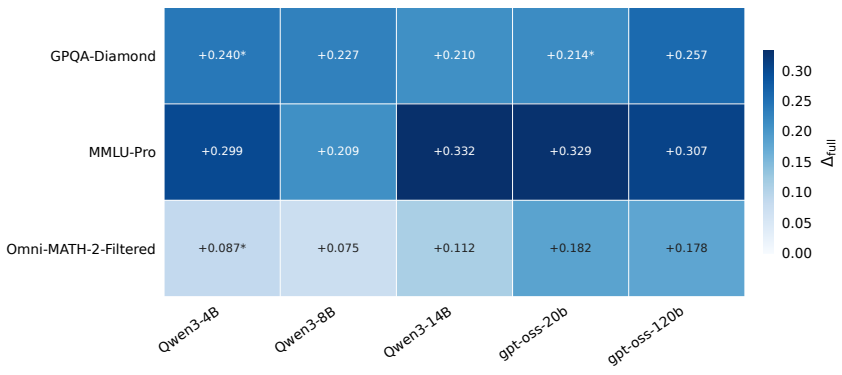
```
<|im_start|>system
Compress the reasoning trace into at most 90 words without solving the problem again. Preserve uncertainty,
      contradictions, candidate answers, and self-corrections. Do not add evidence, verify the answer, repair
      mistakes, or mention information absent from the trace.<|im_end|>
<|im_start|>user
Reasoning trace:
Okay, let's try to figure out which exoplanet has the highest density. The question gives me four options, and
       I need to determine which one has the highest density. Let me start by recalling what density is.
      Density is mass divided by volume, right?

[... full reasoning trace supplied to the summarizer ...]

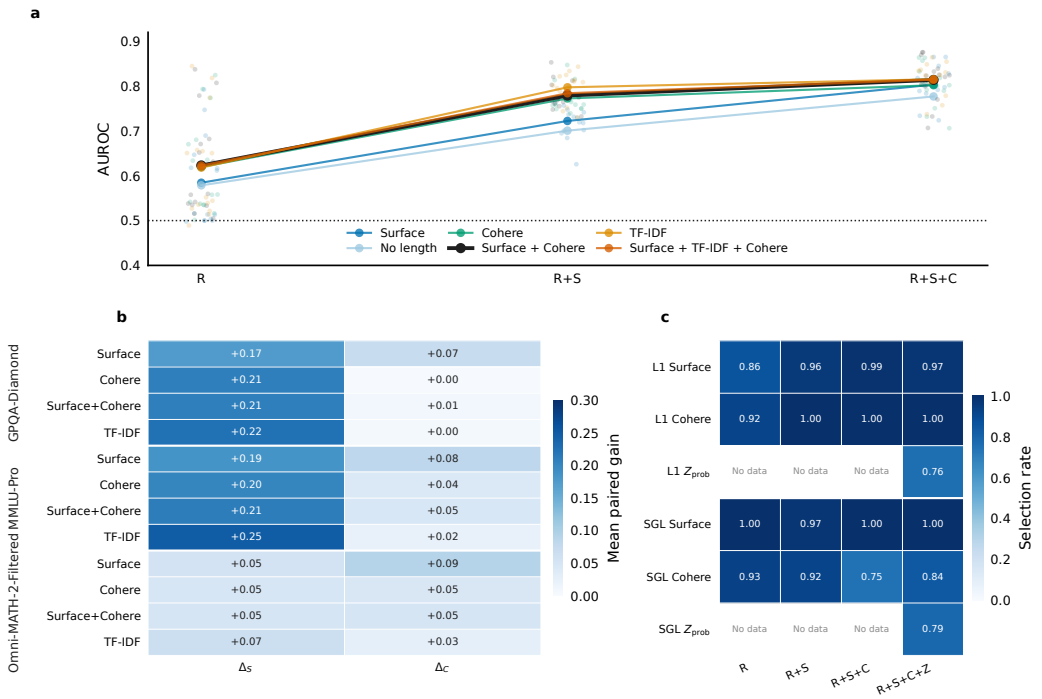
So, in that case, the answer is option c, which is D.

Summary:<|im_end|>
<|im_start|>assistant
The reasoning traces uncertainty about density calculations, noting that same-composition planets with higher
      mass might have higher density due to compression, while others have same density. Contradictions arise
      between scaling assumptions and physical compression effects. Candidate answers include options a, b, c,
       and d, with c potentially having higher density due to compression. Self-corrections suggest
      uncertainty in whether same-composition mass scaling leads to same or higher density.
```

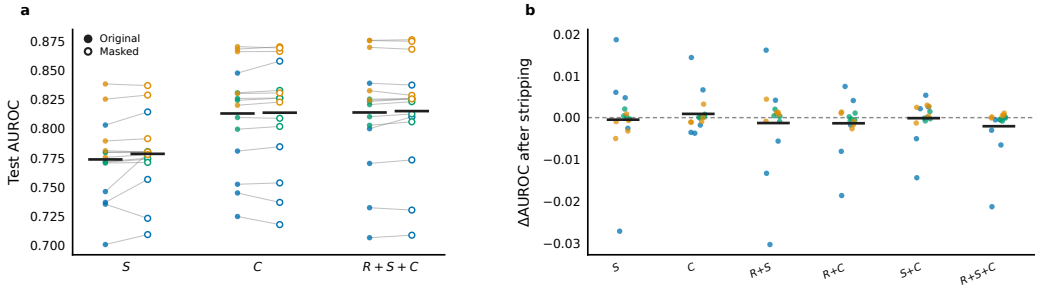
Appendix Figure A7: **Example self-summary prompt and output for the summary level.** The summarizer receives only the reasoning trace C, not the task prompt, gold answer, correctness label, logits, or hidden states. The displayed prompt is shortened. The generated summary is the output for item 1 from Qwen3-4B on GPQA-Diamond, decoded with a 256-token limit under the instruction requesting at most 90 words.



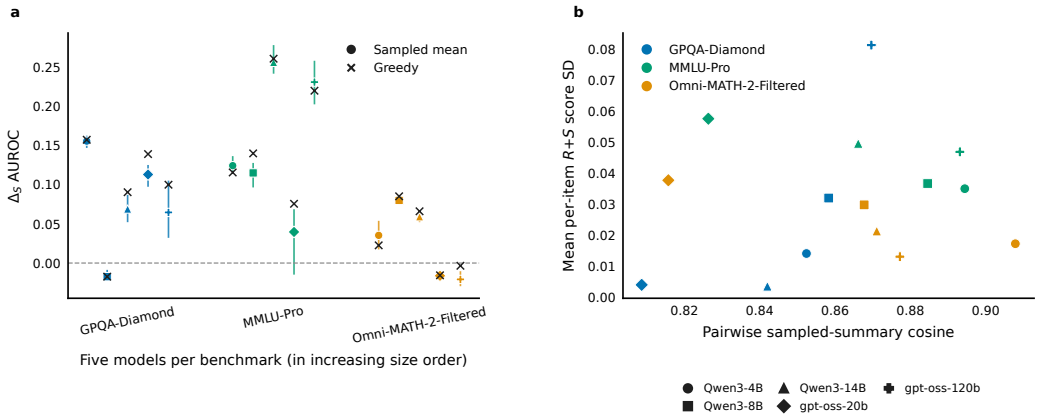
Appendix Figure A8: **Full-ladder gains across the model–benchmark matrix.** Each tile is one target model and one benchmark. Color shows the paired full-ladder gain $\Delta_{\text{full}} = \text{AUROC}(R + S + C + Z) - \text{AUROC}(R)$, also printed on each tile. An asterisk marks settings whose public-level point estimates do not satisfy $R < R + S < R + S + C$.



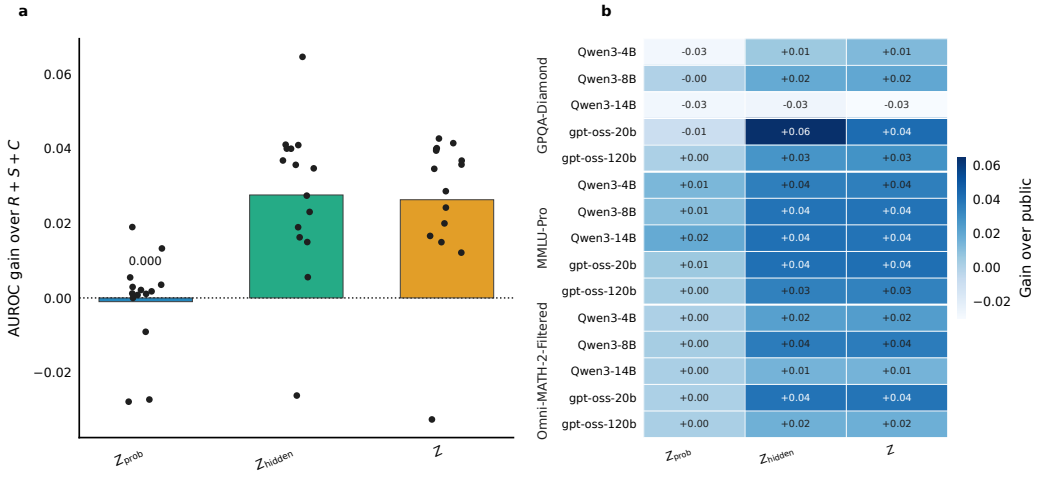
Appendix Figure A9: **The mean public access pattern persists across the tested feature families.** Public monitors use only the text exposed at the corresponding level, never the target model's hidden states, logits, token probabilities, or the original question. **a**, Mean public AUROC increases across access levels for surface features, the no-length version, Cohere embed-v4.0, TF-IDF, and their combinations. Faint points show available model–benchmark settings. All displayed feature families cover the 15 model–benchmark settings. **b**, Per-benchmark paired step gains show that the summary step Δ_S is concentrated on GPQA-Diamond and MMLU-Pro, while Omni-MATH-2-Filtered spreads its public gain more evenly. **c**, Plain LASSO and sparse group LASSO selection rates show that both surface features and embeddings enter the probes. Answer-probability features appear only at the top level and are not part of the public monitor.



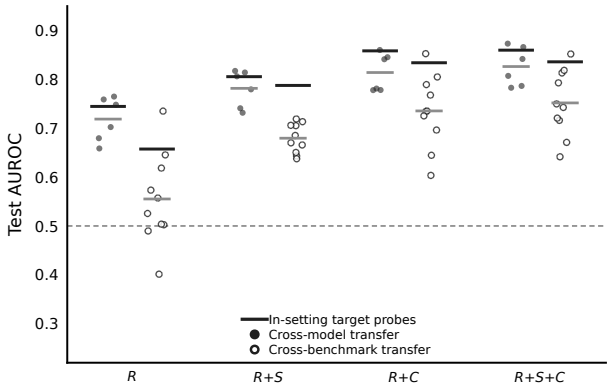
Appendix Figure A10: **Masking and prompt-overlap stripping leave the public signal essentially unchanged.** **a**, Summary, trace, and full public displays after replacing stated answers, option text, and prompt-derived numbers with fixed markers (open circles) match the unmasked originals (filled circles). Gray lines connect the same setting. The combined mask shifts mean trace AUROC by $+0.001$ (Appendix Table A45). **b**, Removing 8–14-word spans copied from the question changes mean AUROC by between -0.002 and $+0.001$ across the six summary- and trace-containing displays (Appendix Table A44).



Appendix Figure A11: **A fixed primary monitor is stable across alternate sampled summaries.** Each point uses a fixed surface-plus-Cohere monitor evaluated on a balanced 48-item subset in one of 15 target-model–benchmark settings. Colors denote benchmarks and marker shapes denote target models. **a**, Colored symbols show the mean equal-stratum summary increment, $\Delta_S = \mathrm{AUROC}(R+S) - \mathrm{AUROC}(R)$, across three sampled summaries, vertical lines span their range, and black crosses show greedy decoding. Across settings, the sampled mean is $+0.086$ AUROC (hierarchical 95% bootstrap interval $[+0.027, +0.146]$), and the greedy value is $+0.010$ higher on average. **b**, For each setting, mean pairwise cosine similarity among the three sampled-summary embeddings is plotted against the mean per-item standard deviation of the fixed $R+S$ monitor score across those summaries.



Appendix Figure A12: **The prompt-withheld replay gain is mostly a hidden-state signal.** Gains are paired AUROC differences relative to the full public display $P = R + S + C$. **a**, The average breakdown compares answer-probability, hidden-state, and combined replay features against the same full public baseline. Bars are means across settings and black dots show setting-level values. The annotated answer-probability bar has a mean gain of 0.000. **b**, The model-by-benchmark heatmap shows where the top-level gain appears for those three replay feature families.



Appendix Figure A13: **Public probes transfer across models and benchmarks.** Transfer tests fit public probes in one setting and evaluate them, without refitting, on held-out settings. Filled circles show cross-model transfer within Omni-MATH-2-Filtered, open circles show cross-benchmark transfer between MMLU-Pro and Omni-MATH-2-Filtered for the same model, black ticks mark matched in-setting target means, and gray ticks mark transfer means. Cross-model transfer loses modest discrimination relative to matched in-setting target probes (0.826 versus 0.860 mean AUROC for $R + S + C$), while cross-benchmark transfer is lower, with mean AUROC 0.752 for $R + S + C$ and individual settings below chance at the shortest displays.

APPENDIX TABLES

The tables below report complete setting-level estimates and supporting diagnostics. They include calibration, secondary metrics, answer and parser baselines, selective risk, failure strata, replay quality, feature-selection checks, and repeated-run comparisons using trace endings (Appendix Tables A30 to A32, A55, A57, A66, A68, A70, A73 and A74).

Appendix Table A1: **Paired uncertainty for the ordered prompt-visible ladder.** Held-out first runs. Entries are AUROC increments and 95% question-bootstrap intervals. Questions are resampled jointly across displays, producers, and overlapping splits within benchmark; fitted scores are held fixed. Each benchmark gives five producer settings equal weight, and the overall mean gives all fifteen settings equal weight. Trace substitution has mean $+0.043$ $[+0.036, +0.050]$ and is reported separately in Table A6.

Population	Δ_S	Δ_C	Δ_Z
GPQA-Diamond	$+0.003$ $[-0.013, +0.019]$	$+0.011$ $[-0.009, +0.032]$	$+0.022$ $[-0.007, +0.050]$
MMLU-Pro	$+0.037$ $[+0.032, +0.041]$	$+0.057$ $[+0.053, +0.063]$	$+0.038$ $[+0.035, +0.042]$
Omni-MATH-2-Filtered	$+0.018$ $[+0.014, +0.022]$	$+0.054$ $[+0.047, +0.060]$	$+0.028$ $[+0.023, +0.033]$
Overall mean	$+0.019$ $[+0.013, +0.025]$	$+0.041$ $[+0.034, +0.048]$	$+0.029$ $[+0.020, +0.039]$

Appendix Table A2: **Trace added after the summary within repeated questions.** Concordance on the complete producer-specific held-out discordant MMLU-Pro samples, with three runs and two correct–incorrect pairs per question. The GPT row uses the ten-display comparison on all 1,389 questions. Each interval jointly resamples questions across displays for 5,000 draws, retaining all runs; it excludes fitting and repeated-scoring variability. This ordered increment differs from trace substitution in the six-display comparison.

Producer	Reader	n	$q+R+S$	$q+R+S+C$	Δ_C [95% interval]
Qwen3-4B	Linear	1,029	0.543	0.584	$+0.041$ $[+0.011, +0.071]$
Qwen3-8B	Linear	969	0.547	0.567	$+0.020$ $[-0.010, +0.050]$
Qwen3-14B	Linear	939	0.486	0.552	$+0.066$ $[+0.034, +0.098]$
gpt-oss-20b	Linear	1,389	0.536	0.560	$+0.025$ $[+0.001, +0.049]$
gpt-oss-120b	Linear	978	0.531	0.564	$+0.033$ $[+0.004, +0.062]$
gpt-oss-20b	GPT-5-mini	1,389	0.776	0.721	-0.055 $[-0.074, -0.036]$

Appendix Table A3: **The prompt-visible trace and replay increments persist under clean-output restrictions.** Held-out first runs with the primary reader. Strict cleaning additionally requires parser status ok, a nonempty response and a nonempty extracted answer. n sums unique held-out producer–question records across settings, so shared questions remain dependent. Each row uses equal setting weights and paired 95% question-bootstrap intervals. Restriction changes the observed population; it is not an intervention repairing faulty outputs.

Restriction	n	Trace after summary	Replay after public
All observed outputs	55,301	$+0.041$ $[+0.034, +0.048]$	$+0.029$ $[+0.020, +0.039]$
Nonempty trace	54,425	$+0.043$ $[+0.036, +0.051]$	$+0.028$ $[+0.019, +0.038]$
Strict clean-output subset	53,837	$+0.044$ $[+0.036, +0.051]$	$+0.028$ $[+0.018, +0.037]$

Appendix Table A4: **Summary access plus trace counts and its full-text extension on held-out first runs.** Lower access is $[\phi(q), \phi(R), \phi(S), M(C)]$ and full-text access appends $\phi(C)$ unchanged. Entries are mean held-out AUROC and paired 95% question-bootstrap intervals. Benchmark means give five producers equal weight; the overall mean gives all fifteen settings equal weight. Metadata transformations, original item splits and fitting settings are fixed before fitting. These are diagnostic comparisons, not a semantic or causal decomposition.

Population	Summary + metadata	Also full trace	Trace increment
GPQA-Diamond	0.740 $[0.686, 0.789]$	0.752 $[0.700, 0.799]$	$+0.011$ $[-0.001, +0.023]$
MMLU-Pro	0.811 $[0.804, 0.818]$	0.821 $[0.814, 0.828]$	$+0.010$ $[+0.008, +0.012]$
Omni-MATH-2-Filtered	0.846 $[0.836, 0.855]$	0.857 $[0.848, 0.866]$	$+0.012$ $[+0.010, +0.014]$
Overall mean	0.799 $[0.781, 0.816]$	0.810 $[0.793, 0.826]$	$+0.011$ $[+0.007, +0.015]$

Appendix Table A5: **The trace-count comparison within repeated questions.** Prompt-visible concordance on producer-specific discordant MMLU-Pro questions. All three runs and the complete lower-level representation are retained. Scores average only fits for which a question was held out. A 5,000-draw question bootstrap retains shared questions across producers, and the overall mean gives the five producers equal weight. Intervals condition on fitted scores. n is the number of questions for a producer; samples overlap across producers.

Producer	n	Summary + metadata	Also full trace	Trace increment
Qwen3-4B	1,029	0.574 $[0.549, 0.599]$	0.588 $[0.561, 0.613]$	$+0.014$ $[-0.002, +0.028]$
Qwen3-8B	969	0.569 $[0.542, 0.596]$	0.571 $[0.543, 0.597]$	$+0.002$ $[-0.015, +0.019]$
Qwen3-14B	939	0.543 $[0.516, 0.569]$	0.549 $[0.522, 0.575]$	$+0.006$ $[-0.009, +0.022]$
gpt-oss-20b	1,389	0.552 $[0.530, 0.574]$	0.565 $[0.543, 0.587]$	$+0.013$ $[-0.001, +0.027]$
gpt-oss-120b	978	0.553 $[0.526, 0.580]$	0.559 $[0.533, 0.586]$	$+0.006$ $[-0.011, +0.023]$
Overall mean	–	0.558 $[0.547, 0.570]$	0.566 $[0.555, 0.579]$	$+0.008$ $[+0.001, +0.015]$

Appendix Table A6: **Direct display contrasts on held-out first runs.** Fifteen-setting overall means with paired 95% question-bootstrap intervals. Shared questions and overlapping splits stay paired.

Comparison	Estimate	95% interval
Summary added, no prompt	$+0.156$	$[+0.133, +0.179]$
Trace versus summary, no prompt	$+0.033$	$[+0.023, +0.042]$
Summary added, prompt visible	$+0.019$	$[+0.013, +0.025]$
Trace versus summary, prompt visible	$+0.043$	$[+0.036, +0.050]$
Prompt difference in summary increment	-0.137	$[-0.159, -0.113]$

Appendix Table A7: **The reader interaction is estimated directly.** Within-question concordance differences on the same 1,389 discordant gpt-oss-20b questions. All displays include the prompt unless labelled otherwise. Summary and trace additions each use $q + R$ as baseline, or R when the prompt is withheld; neither retains the other reasoning display. Intervals resample questions jointly across readers and displays.

Reader	Contrast	Estimate	95% interval
Linear	Summary added	$+0.051$	$[+0.019, +0.083]$
Linear	Trace versus summary	$+0.018$	$[-0.008, +0.045]$
Linear	Trace after response	$+0.069$	$[+0.037, +0.102]$
Linear	Summary added, no prompt	-0.014	$[-0.045, +0.018]$
Linear	Trace versus summary, no prompt	$+0.044$	$[+0.018, +0.071]$
GPT-5-mini	Summary added	-0.024	$[-0.040, -0.007]$
GPT-5-mini	Trace versus summary	-0.040	$[-0.059, -0.022]$
GPT-5-mini	Trace after response	-0.064	$[-0.084, -0.044]$
GPT-5-mini	Summary added, no prompt	$+0.153$	$[+0.126, +0.181]$
GPT-5-mini	Trace versus summary, no prompt	$+0.034$	$[+0.010, +0.058]$
GPT-5-mini minus Linear	Summary added	-0.075	$[-0.110, -0.038]$
GPT-5-mini minus Linear	Trace versus summary	-0.058	$[-0.089, -0.027]$
GPT-5-mini minus Linear	Trace after response	-0.133	$[-0.171, -0.096]$
GPT-5-mini minus Linear	Summary added, no prompt	$+0.167$	$[+0.127, +0.207]$
GPT-5-mini minus Linear	Trace versus summary, no prompt	-0.010	$[-0.045, +0.024]$

Appendix Table A8: **Selection accuracy on each producer's complete held-out question union.** Percent correct after selecting one of three completed runs. Concordant questions contribute their known common outcome. The three monitoring displays all show the prompt. Majority means the plurality extracted answer, with lowest run identifier breaking ties; the same tie policy applies to reader scores. Producer-specific populations differ.

Producer	Reader	Questions	Random	Majority	$q + R$	$q + R + S$	$q + R + C$	Oracle
Qwen3-4B	Linear	8,126	71.06	71.57	71.01	71.61	72.02	77.34
Qwen3-8B	Linear	8,057	75.33	76.12	75.54	75.81	75.92	81.25
Qwen3-14B	Linear	8,102	77.68	78.51	77.46	77.61	78.14	83.36
gpt-oss-20b	Linear	8,123	74.66	76.24	73.73	75.29	75.59	82.94
gpt-oss-20b	GPT-5-mini	8,123	74.66	76.24	79.34	78.84	78.49	82.94
gpt-oss-120b	Linear	8,108	79.12	80.08	78.52	79.46	79.87	84.95

Appendix Table A9: **Display usefulness depends on prompt visibility and reader.** Within-question concordance on the same outcome-discordant MMLU-Pro questions, with three runs per question. Each trace display excludes the summary. Figure A1 shows uncertainty. The machine-readable results contain every interval.

Producer	Reader	n	R	$R + S$	$R + C$	$q + R$	$q + R + S$	$q + R + C$
gpt-oss-20b	Linear	800	0.539	0.510	0.540	0.491	0.521	0.545
gpt-oss-20b	GPT-5-mini	800	0.503	0.656	0.684	0.794	0.774	0.724
gpt-oss-20b	Qwen3.5-27B	800	0.478	0.671	0.751	0.853	0.825	0.796
Qwen3-14B	Linear	800	0.496	0.504	0.559	0.513	0.493	0.558
Qwen3-14B	GPT-5-mini	800	0.518	0.597	0.701	0.735	0.733	0.725
Qwen3-14B	Qwen3.5-27B	800	0.508	0.633	0.738	0.814	0.794	0.790

Appendix Table A10: **All four paired reader interactions are below zero.** Prompt-visible concordance differences on 800 questions from each producer. Summary and trace increments use $q + R$ as baseline. The interaction subtracts the linear trace-minus-summary contrast from the corresponding LLM contrast. Separate 95% question-bootstrap intervals pair all readers, displays, and runs within each producer. Intervals are reported separately for each contrast.

Reader	Contrast	Estimate	95% interval
gpt-oss-20b			
Linear	Summary added	$+0.030$	$[-0.014, +0.074]$
Linear	Trace versus summary	$+0.024$	$[-0.011, +0.060]$
Linear	Trace added	$+0.054$	$[+0.011, +0.096]$
GPT-5-mini	Summary added	-0.021	$[-0.043, +0.001]$
GPT-5-mini	Trace versus summary	-0.049	$[-0.074, -0.025]$
GPT-5-mini	Trace added	-0.070	$[-0.096, -0.045]$
Qwen3.5-27B	Summary added	-0.027	$[-0.043, -0.010]$
Qwen3.5-27B	Trace versus summary	-0.030	$[-0.049, -0.010]$
Qwen3.5-27B	Trace added	-0.057	$[-0.077, -0.038]$
GPT-5-mini minus linear	Interaction I_{C-S}	-0.074	$[-0.116, -0.034]$
Qwen3.5-27B minus linear	Interaction I_{C-S}	-0.054	$[-0.094, -0.014]$
Qwen3-14B			
Linear	Summary added	-0.019	$[-0.064, +0.023]$
Linear	Trace versus summary	$+0.064$	$[+0.028, +0.102]$
Linear	Trace added	$+0.045$	$[+0.001, +0.089]$
GPT-5-mini	Summary added	-0.002	$[-0.025, +0.021]$
GPT-5-mini	Trace versus summary	-0.007	$[-0.031, +0.016]$
GPT-5-mini	Trace added	-0.009	$[-0.033, +0.016]$
Qwen3.5-27B	Summary added	-0.020	$[-0.040, +0.000]$
Qwen3.5-27B	Trace versus summary	-0.004	$[-0.024, +0.016]$
Qwen3.5-27B	Trace added	-0.024	$[-0.046, -0.004]$
GPT-5-mini minus linear	Interaction I_{C-S}	-0.072	$[-0.115, -0.029]$
Qwen3.5-27B minus linear	Interaction I_{C-S}	-0.069	$[-0.112, -0.027]$

Appendix Table A11: **Actual inference use includes invalid attempts and retries.** Qwen3.5-27B counts cover both producers. GPT-5-mini counts cover Qwen3-14B only. Each producer sample contains 800 questions. Pilot use is excluded. GPT-on-gpt-oss use is reported separately in Appendix A.2. Output tokens include reasoning. Token counts use each reader's tokenizer and do not establish compute matching.

Reader	Output allowance	Attempts	Valid added	Input tokens	Output tokens
Qwen3.5-27B	32,768 / 65,536	28,833	28,799	36,969,423	99,416,512
Qwen3.5-27B	65,536 (exception)	1	1	124	3,481
GPT-5-mini	512	14,400	4,943	22,315,552	6,723,538
GPT-5-mini	8,192	9,457	9,440	17,405,573	11,450,181
GPT-5-mini	Default	17	17	45,744	99,260

Qwen3.5-27B retries invalid judgments at a larger output allowance with unchanged prompt and sampling settings, and every attempt is retained. GPT retries only invalid judgments, keeping the same instruction and every first-valid score. Default means no explicit API output limit. Total GPT-on-Qwen use is 39,766,869 input and 18,272,979 output tokens, including 17,685,504 reasoning tokens.

Appendix Table A12: **Notation for the observability ladder.** Components are derived from one fixed completed run per item. Levels are the displays a monitor receives.

Symbol	Object	Definition
q	Task prompt	Benchmark question text; visible to prompt-conditioned monitors.
R	Final response	The model's final response text from the completed run.
S	Self-summary	Post-hoc self-summary of the trace, generated from C only by the same target model under an instruction requesting at most 90 words.
C	Reasoning trace	The full reasoning trace of the completed run.
Z	Provider-only replay features	Answer-probability (Z_{prob}) and hidden-state (Z_{hidden}) features extracted from a prompt-conditioned replay of the target model on q and C.
$\mathcal{O}_1^q$	Prompt-visible level 1	$q + R$
$\mathcal{O}_2^q$	Prompt-visible level 2	$q + R + S$
$\mathcal{O}_3^q$	Prompt-visible level 3	$q + R + S + C$ (full prompt-visible public display)
$\mathcal{O}_4^q$	Provider-only features	$q + R + S + C + Z$ (public display plus replay-derived features)

Appendix Table A13: **The model set gives a within-family scale sweep and a second-family contrast.** All target models are treated as locally inspectable open-weight systems for the provider-only replay features. Public levels restrict the monitor input to model-produced visible text. The final level adds hidden states and token probabilities from a prompt-conditioned replay of the target model. gpt-oss parameter counts are total parameters, with active parameters shown in parentheses. gpt-oss inference uses the model-specific Transformers code path and locally stored Hugging Face kernel components required by the released weights.

Model	Family	Size	Hidden states	Token probabilities	Context window
Qwen/Qwen3-4B	Qwen3	4B	Full	Full	32,768
Qwen/Qwen3-8B	Qwen3	8B	Full	Full	32,768
Qwen/Qwen3-14B	Qwen3	14B	Full	Full	32,768
openai/gpt-oss-20b	gpt-oss	21B (3.6B active)	Full	Full	128,000
openai/gpt-oss-120b	gpt-oss	117B (5.1B active)	Full	Full	128,000

Appendix Table A14: **Settings used to generate reasoning traces and summaries and to calculate model-based measurements.** The reported analysis can be reproduced from the saved outputs even when the original text generation was stochastic. Reasoning traces were generated with vLLM's default sampling settings except for the token limit. Summaries were generated with Hugging Face `generate` using each loaded model's generation settings, with no explicit temperature or top-p override.

Component	Temperature	Top-p	Top-k	Random seed	Max tokens	Stopping behavior	Deterministic?
Trace generation (vLLM chat template)	1.0	1.0	0	None	Qwen: 28,000; gpt-oss: 126,000	no added stopping rule; model EOS and chat-template stopping apply	No; the analysis uses saved traces
Summary generation (HF, `do_sample=True`, `use_cache=False`)	model setting; default 1.0 if unset	model setting; default 1.0 if unset	model setting; default 50 if unset	seed identifier 0 recorded; random-number generator not fixed	256 minus any final-channel prefill	EOS used for padding; summaries were not shortened afterward	No; the analysis uses saved summaries
Calculation of Z (forward pass only)	not applicable; no sampling	not applicable; no sampling	not applicable; no sampling	not applicable; deterministic forward pass	not applicable; no new text generated	model-specific final suffixes; bare-label and answer-span scores taken at the final position	Yes, apart from numeric differences across hardware and libraries

Appendix Table A15: **The benchmark set includes multiple-choice and open-response tasks.** GPQA-Diamond and MMLU-Pro use answer-choice log-probability features in Z_{prob}. Omni-MATH-2-Filtered uses answer-span log-probability features for the model's own extracted final answer, including answer-token and answer-character counts and the sum of token log probabilities.

Benchmark	Domain	Size	Answer type	Verification and Z_{prob}
GPQA-Diamond	Graduate science	198	Four-way multiple choice	Option match; log-probabilities over A–D
MMLU-Pro	Multi-task language understanding	12,032	Up to ten-way multiple choice	Option match; log probabilities for fixed choices A–J, with unavailable choices masked out
Omni-MATH-2-Filtered	Olympiad mathematics	4,181	Open-response numeric / expression	Stored GPT-5-mini judgment of whether the model response matches the reference answer; answer-span log-probability features for the model's extracted final answer, including answer-token and answer-character counts and the sum of token log probabilities

Appendix Table A16: **Correctness prevalence varies by benchmark and target model.** The table reports the binary final-answer correctness labels used in the main ladder analyses. GPQA-Diamond and MMLU-Pro use benchmark option matching. Omni-MATH-2-Filtered correctness is determined by GPT-5-mini equivalence judgments. AUROC is not mechanically shifted by these prevalences because it ranks correct against incorrect runs, but prevalence affects AUPRC baselines, calibration metrics, and uncertainty.

Benchmark	Target model	n	Correct	Correctness rate
GPQA-Diamond	Qwen3-4B	198	110	55.6%
GPQA-Diamond	Qwen3-8B	198	115	58.1%
GPQA-Diamond	Qwen3-14B	198	124	62.6%
GPQA-Diamond	gpt-oss-20b	198	124	62.6%
GPQA-Diamond	gpt-oss-120b	198	138	69.7%
MMLU-Pro	Qwen3-4B	12,032	8,550	71.1%
MMLU-Pro	Qwen3-8B	12,032	9,073	75.4%
MMLU-Pro	Qwen3-14B	12,032	9,375	77.9%
MMLU-Pro	gpt-oss-20b	12,032	8,945	74.3%
MMLU-Pro	gpt-oss-120b	12,032	9,555	79.4%
Omni-MATH-2-Filtered	Qwen3-4B	4,181	2,751	65.8%
Omni-MATH-2-Filtered	Qwen3-8B	4,181	2,455	58.7%
Omni-MATH-2-Filtered	Qwen3-14B	4,181	2,853	68.2%
Omni-MATH-2-Filtered	gpt-oss-20b	4,181	3,016	72.1%
Omni-MATH-2-Filtered	gpt-oss-120b	4,181	3,131	74.9%

Appendix Table A17: **Public-text embeddings are not provider-side replay features.** Public monitors receive only the model-produced text exposed at the corresponding access level. Replay features are extracted from a prompt-conditioned forward pass of the target model on the task prompt and reasoning trace. We record answer-extraction and replay outcomes separately so that failures can be examined in sensitivity checks.

Feature family	Input to feature extractor	Uses provider-side replay?	Level
Surface text	Visible public text P	no	public
No-length surface text	Visible public text P with direct length/count features removed	no	public
TF-IDF text	Visible public text P	no	public
Cohere embed-v4.0	Visible public text P	no	public
Task prompt q_i	Benchmark prompt text, rendered into the display	no	control levels only
Target-model answer log-probabilities	$(q_i, C_i, s_{\text{stop}})$ through target model; open-response spans include length-sensitive count and sum features	yes	Z_{prob}
Target-model raw hidden state	$(q_i, C_i, s_{\text{stop}})$ through target model	yes	Z_{hidden}
Answer extraction and replay checks	Extraction outcome, replay outcome, and replay context-fit indicator	no	Checks

Appendix Table A18: **Primary public surface features use rates for lexical markers.** Features are computed only from the public text visible at the corresponding access level. Raw marker counts and additional length/log-count variants are retained as appendix diagnostics, but the primary surface family uses the rate-based definitions below.

Feature	Definition	Intended signal
Word count	Number of alphabetic word matches in the visible text	Overall trace length
Sentence count	Number of sentence-ending punctuation markers	Reasoning segmentation
Average word length	Mean character length of alphabetic words	Lexical form and technical vocabulary
Digit density	Digit characters divided by total characters	Numerical reasoning content
Math-character density	Math-like characters divided by total characters	Formulaic or symbolic content
Uncertainty rate	Uncertainty markers divided by word count	Explicit hesitation or low-confidence language
Self-correction rate	Self-correction markers divided by word count	Revisions, reversals, and reconsideration
Answer-marker rate	Answer markers divided by word count	Explicit conclusion or answer-framing language

Appendix Table A19: **Inputs used by each correctness probe.** All probes predict the same binary target, final-answer correctness. The reference answer and correctness label are never included as inputs. They are used only to define y_i. Prompt-withheld public levels exclude the task prompt as a separate input and exclude provider-side replay features. The Z levels add features extracted from a prompt-conditioned replay of the task prompt and reasoning trace. Across the reported comparisons, the provider-side Z features appear only alongside the full public display; they are never used as a public display or monitor by themselves. The primary hidden-state model uses the raw final hidden state with ℓ_2 shrinkage.

Probe	Feature vector in the logistic model	Information not used
Response only (R)	$\boldsymbol{z}_i = \phi_f(R_i)$	Task prompt q_i, self-summary S_i, reasoning trace C_i, replay features Z_i, reference answer, correctness label
Response + summary ($R + S$)	$\boldsymbol{z}_i = \phi_f(R_i \Vert S_i)$	Task prompt q_i, full reasoning trace C_i, replay features Z_i, reference answer, correctness label
Response + summary + trace ($R + S + C$)	$\boldsymbol{z}_i = \phi_f(R_i \Vert S_i \Vert C_i)$	Task prompt q_i, replay features Z_i, reference answer, correctness label
Public display subsets (S, C, $R + C$, $S + C$)	$\boldsymbol{z}_i = \phi_f(\cdot)$ applied to the corresponding subset of $\{R_i, S_i, C_i\}$	Task prompt q_i, replay features Z_i, reference answer, correctness label; used to compare all component subsets, separately from the main prompt-visible comparison
All prompt-conditioned public combinations (q, $q + R$, $q + S$, $q + C$, $q + R + S$, $q + R + C$, $q + S + C$, $q + R + S + C$)	$\boldsymbol{z}_i = \phi_f(q_i \Vert \cdot)$ applied to the corresponding prompt-conditioned combination	Replay features Z_i, reference answer, correctness label; the primary surface-plus-Cohere analysis follows a fixed sequence of component sets, while the secondary surface-plus-TF-IDF analysis covers all eight combinations
Displays with prompt overlap removed (*noqcopy* variants of S/C-containing levels)	$\boldsymbol{z}_i = \phi_f(\cdot)$ applied to the display after removing 8–14-word spans copied from q_i	Task prompt q_i, replay features Z_i, reference answer, correctness label; tests whether the visible text reconstructs the prompt
Full public + answer probabilities ($R + S + C + Z_{\text{prob}}$)	$\boldsymbol{z}_i = \phi_{\text{all}}(R_i \Vert S_i \Vert C_i) \oplus Z_{\text{prob},i}$	Hidden-state features $Z_{\text{hidden},i}$, reference answer, correctness label
Full public + hidden state ($R + S + C + Z_{\text{hidden}}$)	$\boldsymbol{z}_i = \phi_{\text{all}}(R_i \Vert S_i \Vert C_i) \oplus Z_{\text{hidden},i}$	Answer-probability features $Z_{\text{prob},i}$, reference answer, correctness label
Full public + all replay features ($R + S + C + Z$)	$\boldsymbol{z}_i = \phi_{\text{all}}(R_i \Vert S_i \Vert C_i) \oplus Z_{\text{prob},i} \oplus Z_{\text{hidden},i}$	Reference answer and correctness label
Prompt-conditioned public + replay features ($q + R + S + C + Z_{\text{hidden}}$, $q + R + S + C + Z$)	$\boldsymbol{z}_i = \phi_{\text{all}}(q_i \Vert R_i \Vert S_i \Vert C_i) \oplus Z_i$ or the hidden-state features alone	Reference answer and correctness label; used to separate information in prompt-conditioned public text from information in replay features

Appendix Table A20: **Prompt-withheld observability-ladder AUROC by model–benchmark setting.** Values are mean ± descriptive SD across data splits, not confidence intervals, for the primary public feature family (rate surface plus Cohere) and the provider-side replay feature family (rate surface plus Cohere plus replay-derived features). Omni-MATH-2-Filtered correctness is determined by GPT-5-mini equivalence judgments.

Benchmark	Model	R	$R+S$	$R+S+C$	$R{+}S{+}C{+}Z$	Δ_{full}
GPQA-Diamond	Qwen3-4B	0.505 ± 0.066	0.732 ± 0.063	0.732 ± 0.067	0.744 ± 0.059	+0.240 ± 0.105
GPQA-Diamond	Qwen3-8B	0.631 ± 0.047	0.812 ± 0.082	0.839 ± 0.062	0.859 ± 0.038	+0.227 ± 0.053
GPQA-Diamond	Qwen3-14B	0.558 ± 0.046	0.748 ± 0.056	0.800 ± 0.058	0.767 ± 0.078	+0.210 ± 0.100
GPQA-Diamond	gpt-oss-20b	0.535 ± 0.137	0.755 ± 0.066	0.707 ± 0.096	0.749 ± 0.103	+0.214 ± 0.074
GPQA-Diamond	gpt-oss-120b	0.542 ± 0.073	0.761 ± 0.053	0.770 ± 0.057	0.799 ± 0.069	+0.257 ± 0.123
MMLU-Pro	Qwen3-4B	0.560 ± 0.014	0.764 ± 0.009	0.823 ± 0.010	0.859 ± 0.008	+0.299 ± 0.017
MMLU-Pro	Qwen3-8B	0.652 ± 0.014	0.774 ± 0.013	0.821 ± 0.004	0.861 ± 0.007	+0.209 ± 0.011
MMLU-Pro	Qwen3-14B	0.533 ± 0.004	0.775 ± 0.010	0.825 ± 0.007	0.865 ± 0.003	+0.332 ± 0.007
MMLU-Pro	gpt-oss-20b	0.515 ± 0.009	0.767 ± 0.016	0.803 ± 0.015	0.844 ± 0.011	+0.329 ± 0.005
MMLU-Pro	gpt-oss-120b	0.538 ± 0.006	0.769 ± 0.004	0.810 ± 0.005	0.845 ± 0.005	+0.307 ± 0.009
Omni-MATH-2-Filtered	Qwen3-4B	0.807 ± 0.015	0.803 ± 0.018	0.870 ± 0.015	0.894 ± 0.013	+0.087 ± 0.009
Omni-MATH-2-Filtered	Qwen3-8B	0.837 ± 0.027	0.853 ± 0.026	0.875 ± 0.022	0.912 ± 0.012	+0.075 ± 0.017
Omni-MATH-2-Filtered	Qwen3-14B	0.779 ± 0.010	0.821 ± 0.006	0.876 ± 0.008	0.891 ± 0.005	+0.112 ± 0.015
Omni-MATH-2-Filtered	gpt-oss-20b	0.682 ± 0.019	0.775 ± 0.013	0.824 ± 0.022	0.864 ± 0.020	+0.182 ± 0.026
Omni-MATH-2-Filtered	gpt-oss-120b	0.671 ± 0.012	0.776 ± 0.022	0.832 ± 0.019	0.849 ± 0.011	+0.178 ± 0.010

Appendix Table A21: **Paired AUROC increments along the observability ladder.** Each increment is computed within each data split before averaging. The ± values are descriptive SDs across data splits, not confidence intervals. Descriptive setting-bootstrap intervals over the observed design are reported separately in Appendix Table A22. Omni-MATH-2-Filtered correctness is determined by GPT-5-mini equivalence judgments.

Benchmark	Model	Δ_S	Δ_C	Δ_Z	Δ_{full}
GPQA-Diamond	Qwen3-4B	+0.228 ± 0.106	+0.000 ± 0.038	+0.012 ± 0.023	+0.240 ± 0.105
GPQA-Diamond	Qwen3-8B	+0.181 ± 0.055	+0.027 ± 0.023	+0.020 ± 0.051	+0.227 ± 0.053
GPQA-Diamond	Qwen3-14B	+0.190 ± 0.057	+0.052 ± 0.033	−0.033 ± 0.051	+0.210 ± 0.100
GPQA-Diamond	gpt-oss-20b	+0.220 ± 0.123	−0.049 ± 0.082	+0.043 ± 0.125	+0.214 ± 0.074
GPQA-Diamond	gpt-oss-120b	+0.219 ± 0.117	+0.010 ± 0.027	+0.029 ± 0.026	+0.257 ± 0.123
MMLU-Pro	Qwen3-4B	+0.205 ± 0.019	+0.059 ± 0.002	+0.036 ± 0.003	+0.299 ± 0.017
MMLU-Pro	Qwen3-8B	+0.122 ± 0.006	+0.046 ± 0.012	+0.040 ± 0.003	+0.209 ± 0.011
MMLU-Pro	Qwen3-14B	+0.241 ± 0.010	+0.051 ± 0.009	+0.040 ± 0.006	+0.332 ± 0.007
MMLU-Pro	gpt-oss-20b	+0.251 ± 0.010	+0.036 ± 0.008	+0.041 ± 0.005	+0.329 ± 0.005
MMLU-Pro	gpt-oss-120b	+0.231 ± 0.009	+0.042 ± 0.003	+0.035 ± 0.003	+0.307 ± 0.009
Omni-MATH-2-Filtered	Qwen3-4B	−0.004 ± 0.008	+0.067 ± 0.004	+0.024 ± 0.005	+0.087 ± 0.009
Omni-MATH-2-Filtered	Qwen3-8B	+0.016 ± 0.007	+0.022 ± 0.005	+0.037 ± 0.013	+0.075 ± 0.017
Omni-MATH-2-Filtered	Qwen3-14B	+0.042 ± 0.008	+0.055 ± 0.006	+0.015 ± 0.010	+0.112 ± 0.015
Omni-MATH-2-Filtered	gpt-oss-20b	+0.093 ± 0.019	+0.049 ± 0.013	+0.039 ± 0.021	+0.182 ± 0.026
Omni-MATH-2-Filtered	gpt-oss-120b	+0.106 ± 0.015	+0.056 ± 0.015	+0.017 ± 0.013	+0.178 ± 0.010

Appendix Table A22: **Setting-level resampling describes the prompt-visible and prompt-withheld ladder contrasts.** Mean deltas and descriptive SDs are computed across the 15 model–benchmark settings after averaging the five data splits within each setting. Intervals are descriptive nonparametric 95% bootstrap intervals over the 15 observed settings, not population-generalization intervals. The p value is a one-sided paired sign-flip test of the setting-level mean against zero in the tested direction: positive for ladder increments and negative for prompt-visible-minus-prompt-withheld differences. Values within 10^{-12} AUROC of zero are treated as ties.

Condition	Quantity	Contrast	Mean	Between-setting SD	95% bootstrap interval	Settings in tested direction	p
Prompt-visible	$q\Delta_S$	$q{+}R$ to $q{+}R{+}S$	+0.019	0.017	[+0.011, +0.027]	11/15	$<$ 0.001
Prompt-visible, trace first	$q\Delta_{C\|R}$	$q{+}R$ to $q{+}R{+}C$	+0.062	0.033	[+0.046, +0.077]	15/15	$<$ 0.001
Prompt-visible	$q\Delta_C$	$q{+}R{+}S$ to $q{+}R{+}S{+}C$	+0.041	0.026	[+0.028, +0.052]	14/15	$<$ 0.001
Prompt-visible, trace first	$q\Delta_{S\|R,C}$	$q{+}R{+}C$ to $q{+}R{+}S{+}C$	−0.002	0.009	[−0.007, +0.002]	8/15	0.819
Prompt-visible direct comparison	$q(C{-}S)$	$q{+}R{+}S$ versus $q{+}R{+}C$	+0.043	0.020	[+0.033, +0.052]	15/15	$<$ 0.001
Prompt-visible	$q\Delta_Z$	$q{+}R{+}S{+}C$ to $q{+}R{+}S{+}C{+}Z$	+0.029	0.014	[+0.023, +0.036]	15/15	$<$ 0.001
Prompt-visible	$q\Delta_{\text{public}}$	$q{+}R$ to $q{+}R{+}S{+}C$	+0.060	0.039	[+0.040, +0.077]	14/15	$<$ 0.001
Prompt-visible	$q\Delta_{\text{full}}$	$q{+}R$ to $q{+}R{+}S{+}C{+}Z$	+0.089	0.045	[+0.066, +0.110]	15/15	$<$ 0.001
Prompt-withheld	Δ_S	R to $R{+}S$	+0.156	0.087	[+0.113, +0.196]	14/15	$<$ 0.001
Prompt-withheld	Δ_C	$R{+}S$ to $R{+}S{+}C$	+0.035	0.030	[+0.019, +0.048]	13/15 (+1 tie)	$<$ 0.001
Prompt-withheld	Δ_Z	$R{+}S{+}C$ to $R{+}S{+}C{+}Z$	+0.026	0.019	[+0.016, +0.034]	14/15	$<$ 0.001
Prompt-withheld	Δ_{public}	R to $R{+}S{+}C$	+0.191	0.080	[+0.151, +0.228]	15/15	$<$ 0.001
Prompt-withheld	Δ_{full}	R to $R{+}S{+}C{+}Z$	+0.217	0.082	[+0.177, +0.257]	15/15	$<$ 0.001
Paired condition difference	$q\Delta_S - \Delta_S$	Prompt-visible minus prompt-withheld summary increment	−0.137	0.086	[−0.177, −0.095]	13/15	$<$ 0.001
Paired condition difference	$q\Delta_C - \Delta_C$	Prompt-visible minus prompt-withheld trace increment	+0.006	0.015	[−0.002, +0.013]	2/15	0.920
Paired condition difference	$q\Delta_Z - \Delta_Z$	Prompt-visible minus prompt-withheld replay increment	+0.003	0.011	[−0.002, +0.009]	6/15	0.855
Paired condition difference	$q\Delta_{\text{public}} - \Delta_{\text{public}}$	Prompt-visible minus prompt-withheld public gain	−0.131	0.084	[−0.170, −0.089]	13/15	$<$ 0.001
Paired condition difference	$q\Delta_{\text{full}} - \Delta_{\text{full}}$	Prompt-visible minus prompt-withheld full gain	−0.128	0.082	[−0.166, −0.087]	13/15	$<$ 0.001

Appendix Table A23: **The summary adds much less when the prompt is visible.** Values are mean paired AUROC changes across the 15 model–benchmark settings and five data splits (± SD across the 75 combinations). Prompt-visible quantities include both orders in which sources are added and the direct $q{+}R{+}C$ versus $q{+}R{+}S$ comparison. The provider-only comparison adds the narrow prompt-conditioned last-token replay features Z. The prompt-withheld column gives the corresponding output-only comparison. Brackets are descriptive 95% setting-bootstrap intervals after averaging the five data splits within each setting.

Quantity	Prompt-visible	Prompt-withheld	Difference
Summary added first	+0.019 ± 0.022	+0.156 ± 0.099	−0.137 [−0.177, −0.095]
Trace added after summary	+0.041 ± 0.031	+0.035 ± 0.038	+0.006 [−0.002, +0.013]
Trace added first	+0.062 ± 0.036 [+0.046, +0.077]	—	—
Summary added after trace	−0.002 ± 0.020 [−0.007, +0.002]	—	—
Trace minus summary	+0.043 ± 0.026 [+0.033, +0.052]	—	—
Provider-only replay features	+0.029 ± 0.035	+0.026 ± 0.040	+0.003 [−0.002, +0.009]
Public total	+0.060 ± 0.043	+0.191 ± 0.100	−0.131 [−0.170, −0.089]
Full total	+0.089 ± 0.054	+0.217 ± 0.094	−0.128 [−0.166, −0.087]

Appendix Table A24: **Component-block monitors separate access from concatenation compression.** The primary public monitor embeds each rendered display once. The component-block monitor instead concatenates separately computed surface and Cohere feature blocks for the displayed components, so higher-access levels retain the lower-level representation unchanged. Values are mean AUROC across model–benchmark settings after averaging over five data splits (± between-setting SD). Δ is component-block minus primary-display AUROC.

Display	Primary display	Component blocks	Δ	Settings
R	0.623 ± 0.112	0.624 ± 0.112	+0.001 ± 0.002	15
$R{+}S$	0.779 ± 0.031	0.775 ± 0.059	−0.004 ± 0.037	15
$R{+}S{+}C$	0.814 ± 0.048	0.810 ± 0.058	−0.004 ± 0.019	15
q	0.734 ± 0.032	0.735 ± 0.032	+0.001 ± 0.001	15
$q{+}R$	0.750 ± 0.042	0.748 ± 0.051	−0.001 ± 0.017	15
$q{+}R{+}S$	0.769 ± 0.038	0.784 ± 0.049	+0.016 ± 0.021	15
$q{+}C$	0.812 ± 0.045	0.815 ± 0.051	+0.003 ± 0.013	15
$q{+}R{+}S{+}C$	0.809 ± 0.050	0.812 ± 0.056	+0.002 ± 0.015	15

Appendix Table A25: **Summary–trace and prompt-visible conclusions persist across the observed design.** Mean paired AUROC contrasts shown separately by benchmark, recomputed after excluding each model family or benchmark, and under item-count weighting instead of equal weighting. The excess ratio is $(\mathrm{AUROC}(S) - 0.5)/(\mathrm{AUROC}(C) - 0.5)$; it compares excess AUROC rather than retained information. Its paired setting-bootstrap estimate is 0.874 (95% descriptive interval [0.855, 0.896]; 2,000 draws over the 15 observed settings).

Aggregate	Δ_{full}	$C - S$	$(R{+}C) - (R{+}S)$	$q\Delta_S$	$q\Delta_C$	$(q{+}R{+}C) - (q{+}R{+}S)$	$q\Delta_{\mathrm{public}}$	Excess ratio
All settings (equal weight)	+0.217	+0.039	+0.033	+0.019	+0.041	+0.043	+0.060	0.874
Benchmark only GPQA-Diamond	+0.230	+0.026	+0.006	+0.003	+0.011	+0.021	+0.014	0.905
Benchmark only MMLU-Pro	+0.295	+0.043	+0.046	+0.037	+0.057	+0.056	+0.094	0.864
Benchmark only Omni-MATH-2-Filtered	+0.127	+0.049	+0.047	+0.018	+0.054	+0.052	+0.071	0.860
Excluding family gpt-oss	+0.199	+0.043	+0.041	+0.017	+0.041	+0.044	+0.059	0.868
Excluding family Qwen3	+0.245	+0.033	+0.021	+0.021	+0.040	+0.042	+0.061	0.885
Excluding benchmark GPQA-Diamond	+0.211	+0.046	+0.046	+0.027	+0.056	+0.054	+0.083	0.862
Excluding benchmark MMLU-Pro	+0.178	+0.037	+0.026	+0.010	+0.032	+0.036	+0.043	0.880
Excluding benchmark Omni-MATH-2-Filtered	+0.262	+0.035	+0.026	+0.020	+0.034	+0.039	+0.054	0.883
All settings (item weighted)	+0.251	+0.045	+0.046	+0.031	+0.056	+0.054	+0.087	0.863

Appendix Table A26: **Item-count weighting preserves the main output-only and prompt-visible increments.** Equal weighting assigns one vote to each of the 15 model–benchmark settings. Item weighting instead weights each setting by its mean held-out test count across the five data splits. The final column is the item-weighted estimate minus the equal-setting estimate.

Increment	Equal-weight mean	Item-weighted mean	Change
Δ_S	+0.156	+0.169	+0.013
Δ_C	+0.035	+0.047	+0.012
Δ_Z	+0.026	+0.035	+0.009
Δ_{public}	+0.191	+0.216	+0.026
Δ_{full}	+0.217	+0.251	+0.034
$q\Delta_S$	+0.019	+0.031	+0.012
$q\Delta_C$	+0.041	+0.056	+0.015
$q\Delta_Z$	+0.029	+0.036	+0.006
$q\Delta_{\text{public}}$	+0.060	+0.087	+0.028
$q\Delta_{\text{full}}$	+0.089	+0.123	+0.034

Appendix Table A27: **Mixed-model sensitivity estimates are positive, although the prompt-visible summary interval includes zero.** Intercept-only linear mixed models fit by REML to the 15 per-setting AUROC increments. The preferred specification includes benchmark and model-family variance components; if that fit fails to converge or yields nonfinite fixed-effect uncertainty, the estimate uses the predefined benchmark-only alternative. Estimates and 95% Wald intervals are descriptive model-based sensitivities for the observed design, not population-generalization intervals. Interval endpoints are reported to four decimals to resolve near-zero bounds. Variance components are reported $\times 10^4$; an omitted component is shown as '–'.

Increment	Specification	Estimate	95% interval	σ^2_{bench}	σ^2_{family}	σ^2_{resid}
$q\Delta_S$ (q+R+S vs. q+R)	Benchmark only	+0.019	[−0.0002, +0.0382]	2.65	–	1.12
$q\Delta_C$ (q+R+S+C vs. q+R+S)	Benchmark + family	+0.041	[+0.0042, +0.0772]	9.95	0.02	1.97
$q\Delta_Z$ (q+R+S+C+Z vs. q+R+S+C)	Benchmark + family	+0.030	[+0.0184, +0.0415]	0.41	0.22	1.45
$q\Delta_{\text{public}}$ (q+R+S+C vs. q+R)	Benchmark only	+0.060	[+0.0129, +0.1066]	16.54	–	3.10
$q\Delta_{\text{full}}$ (q+R+S+C+Z vs. q+R)	Benchmark + family	+0.090	[+0.0334, +0.1457]	23.59	0.21	3.40
Δ_S (R+S vs. R)	Benchmark + family	+0.160	[+0.0467, +0.2743]	81.58	11.49	11.16
Δ_C (R+S+C vs. R+S)	Benchmark + family	+0.034	[+0.0042, +0.0635]	4.40	0.91	5.27
Δ_Z (R+S+C+Z vs. R+S+C)	Benchmark + family	+0.027	[+0.0105, +0.0433]	0.88	0.40	2.92
Δ_{public} (R+S+C vs. R)	Benchmark + family	+0.193	[+0.0976, +0.2876]	62.24	2.89	18.34
Δ_{full} (R+S+C+Z vs. R)	Benchmark + family	+0.221	[+0.1164, +0.3254]	69.64	8.61	12.99

Appendix Table A28: **Item-clustered bootstrap intervals for the main ladder.** Held-out test AUROC per setting and level (mean over five data splits) with 95% intervals from resampling test items with replacement within each setting (2,000 draws), so the uncertainty respects item-level clustering of the paired design rather than treating runs as independent.

Benchmark	Model	R	$R+S$	$R+S+C$	$R+S+C+Z$
GPQA-Diamond	Qwen3-4B	0.505 [0.408, 0.604]	0.732 [0.638, 0.817]	0.732 [0.636, 0.822]	0.744 [0.651, 0.835]
GPQA-Diamond	Qwen3-8B	0.631 [0.533, 0.730]	0.812 [0.732, 0.884]	0.839 [0.758, 0.907]	0.859 [0.792, 0.920]
GPQA-Diamond	Qwen3-14B	0.558 [0.452, 0.660]	0.748 [0.655, 0.834]	0.800 [0.716, 0.874]	0.767 [0.675, 0.849]
GPQA-Diamond	gpt-oss-20b	0.535 [0.438, 0.637]	0.755 [0.665, 0.836]	0.707 [0.612, 0.788]	0.749 [0.650, 0.835]
GPQA-Diamond	gpt-oss-120b	0.542 [0.440, 0.642]	0.761 [0.670, 0.843]	0.770 [0.666, 0.858]	0.799 [0.709, 0.876]
MMLU-Pro	Qwen3-4B	0.560 [0.546, 0.574]	0.764 [0.752, 0.777]	0.823 [0.813, 0.833]	0.859 [0.850, 0.868]
MMLU-Pro	Qwen3-8B	0.652 [0.637, 0.667]	0.774 [0.763, 0.786]	0.821 [0.810, 0.831]	0.861 [0.851, 0.870]
MMLU-Pro	Qwen3-14B	0.533 [0.518, 0.549]	0.775 [0.762, 0.787]	0.825 [0.814, 0.836]	0.865 [0.855, 0.875]
MMLU-Pro	gpt-oss-20b	0.515 [0.500, 0.531]	0.767 [0.754, 0.780]	0.803 [0.791, 0.814]	0.844 [0.834, 0.855]
MMLU-Pro	gpt-oss-120b	0.538 [0.521, 0.555]	0.769 [0.755, 0.782]	0.810 [0.798, 0.822]	0.845 [0.834, 0.856]
Omni-MATH-2-Filtered	Qwen3-4B	0.807 [0.789, 0.825]	0.803 [0.785, 0.820]	0.870 [0.855, 0.884]	0.894 [0.879, 0.907]
Omni-MATH-2-Filtered	Qwen3-8B	0.837 [0.821, 0.853]	0.853 [0.838, 0.868]	0.875 [0.861, 0.889]	0.912 [0.899, 0.924]
Omni-MATH-2-Filtered	Qwen3-14B	0.779 [0.758, 0.799]	0.821 [0.803, 0.838]	0.876 [0.861, 0.891]	0.891 [0.877, 0.904]
Omni-MATH-2-Filtered	gpt-oss-20b	0.682 [0.658, 0.706]	0.775 [0.754, 0.795]	0.824 [0.807, 0.841]	0.864 [0.847, 0.879]
Omni-MATH-2-Filtered	gpt-oss-120b	0.671 [0.645, 0.696]	0.776 [0.756, 0.796]	0.832 [0.815, 0.850]	0.849 [0.833, 0.865]

Appendix Table A29: **Label permutation removes the monitor signal at every level.** Held-out test AUROC of the primary probes on real labels (mean over per-setting means $\pm$ between-setting SD) versus probes retrained on label-permuted training data (mean $\pm$ SD over all permutation runs across model–benchmark settings and data splits). Permuted AUROC concentrating at chance confirms chance performance after randomizing the training labels; it is not an independent proof against every form of leakage.

Level	Real labels	Permuted labels	Permutation runs
R	0.623 ± 0.112	0.503 ± 0.062	375
$R+S$	0.779 ± 0.031	0.496 ± 0.057	375
$R+S+C$	0.814 ± 0.048	0.494 ± 0.053	375
$R+S+C+Z$	0.840 ± 0.052	0.496 ± 0.055	375

Appendix Table A30: **AUPRC and Brier generally improve with access, whereas uncalibrated ECE does not.** AUROC, AUPRC, Brier score, and expected calibration error (ECE) for the prompt-withheld ladder levels, averaged over per-setting means (± between-setting SD; per-setting means average five data splits). Lower is better for Brier and ECE. ECE uses 10 equal-width probability bins on the held-out test predictions and is reported for the raw logistic-regression scores without post-hoc calibration.

Benchmark	Level	AUROC	AUPRC	Brier	ECE
GPQA-Diamond	R	0.554 ± 0.047	0.688 ± 0.053	0.249 ± 0.018	0.124 ± 0.018
	$R+S$	0.762 ± 0.030	0.843 ± 0.038	0.226 ± 0.011	0.200 ± 0.028
	$R+S+C$	0.770 ± 0.053	0.842 ± 0.051	0.222 ± 0.016	0.201 ± 0.013
	$R+S+C+Z$	0.784 ± 0.047	0.851 ± 0.057	0.209 ± 0.020	0.204 ± 0.020
MMLU-Pro	R	0.560 ± 0.054	0.793 ± 0.044	0.190 ± 0.019	0.064 ± 0.037
	$R+S$	0.770 ± 0.005	0.907 ± 0.017	0.160 ± 0.011	0.087 ± 0.009
	$R+S+C$	0.816 ± 0.010	0.931 ± 0.011	0.151 ± 0.007	0.096 ± 0.019
	$R+S+C+Z$	0.855 ± 0.010	0.946 ± 0.009	0.143 ± 0.008	0.118 ± 0.038
Omni-MATH-2-Filtered	R	0.755 ± 0.075	0.855 ± 0.028	0.181 ± 0.016	0.107 ± 0.033
	$R+S$	0.806 ± 0.033	0.895 ± 0.011	0.173 ± 0.010	0.115 ± 0.031
	$R+S+C$	0.855 ± 0.025	0.924 ± 0.013	0.154 ± 0.012	0.116 ± 0.036
	$R+S+C+Z$	0.882 ± 0.025	0.937 ± 0.006	0.139 ± 0.018	0.112 ± 0.048
All benchmarks	R	0.623 ± 0.112	0.779 ± 0.082	0.207 ± 0.035	0.099 ± 0.039
	$R+S$	0.779 ± 0.031	0.882 ± 0.037	0.186 ± 0.031	0.134 ± 0.055
	$R+S+C$	0.814 ± 0.048	0.899 ± 0.051	0.176 ± 0.036	0.138 ± 0.053
	$R+S+C+Z$	0.840 ± 0.052	0.911 ± 0.054	0.164 ± 0.037	0.145 ± 0.055

Appendix Table A31: **Platt calibration improves the mean probability estimates while leaving aggregate AUROC essentially unchanged.** Calibrators are fit on validation predictions and applied to held-out test predictions for each benchmark, model, data split, access level, and feature family. Each estimate averages 75 combinations of 15 model–benchmark settings and five data splits. AUROC is shown before calibration; Brier score and expected calibration error (ECE) are shown before and after calibration. An individual fitted calibrator can reverse a ranking when its validation slope is negative.

Level	AUROC	Brier before	Brier after	ECE before	ECE after
q+R	0.750	0.204	0.176	0.153	0.068
q+R+S	0.769	0.190	0.172	0.132	0.068
q+R+S+C	0.809	0.176	0.159	0.130	0.077
q+R+S+C+Z	0.839	0.165	0.147	0.145	0.079
R	0.623	0.207	0.196	0.099	0.050
R+S	0.779	0.186	0.171	0.134	0.077
R+S+C	0.814	0.176	0.158	0.138	0.073
R+S+C+Z	0.840	0.164	0.147	0.145	0.080

Appendix Table A32: **Selected-answer and option-bias baselines do not explain the public monitor signal.** Held-out test AUROC for logistic baselines that see only the selected answer identity, an answer-prior encoding, answer-format indicators, or the selected-answer probability, averaged over five data splits (± SD across data splits). The two rightmost columns repeat the public monitor at R and $R{+}S{+}C$ for reference. Verification fields computed against the gold answer encode the label and are excluded from all baseline inputs.

Benchmark	Model	Answer prior	Selected answer	Answer format	Selected prob.	R monitor	$R{+}S{+}C$ monitor
GPQA-Diamond	Qwen3-4B	0.518 ± 0.072	0.518 ± 0.072	0.510 ± 0.024	0.489 ± 0.137	0.505 ± 0.066	0.732 ± 0.067
GPQA-Diamond	Qwen3-8B	0.544 ± 0.123	0.544 ± 0.123	0.589 ± 0.072	0.576 ± 0.055	0.631 ± 0.047	0.839 ± 0.062
GPQA-Diamond	Qwen3-14B	0.545 ± 0.058	0.545 ± 0.058	0.516 ± 0.017	0.658 ± 0.084	0.558 ± 0.046	0.800 ± 0.058
GPQA-Diamond	gpt-oss-20b	0.526 ± 0.127	0.560 ± 0.146	0.500 ± 0.000	0.625 ± 0.142	0.535 ± 0.137	0.707 ± 0.096
GPQA-Diamond	gpt-oss-120b	0.532 ± 0.063	0.538 ± 0.056	0.500 ± 0.000	0.662 ± 0.109	0.542 ± 0.073	0.770 ± 0.057
MMLU-Pro	Qwen3-4B	0.532 ± 0.018	0.531 ± 0.020	0.507 ± 0.006	0.754 ± 0.011	0.560 ± 0.014	0.823 ± 0.010
MMLU-Pro	Qwen3-8B	0.523 ± 0.016	0.524 ± 0.017	0.557 ± 0.014	0.688 ± 0.015	0.652 ± 0.014	0.821 ± 0.004
MMLU-Pro	Qwen3-14B	0.524 ± 0.010	0.523 ± 0.012	0.508 ± 0.002	0.760 ± 0.007	0.533 ± 0.004	0.825 ± 0.007
MMLU-Pro	gpt-oss-20b	0.512 ± 0.010	0.511 ± 0.011	0.504 ± 0.001	0.652 ± 0.018	0.515 ± 0.009	0.803 ± 0.015
MMLU-Pro	gpt-oss-120b	0.532 ± 0.004	0.522 ± 0.016	0.501 ± 0.001	0.590 ± 0.119	0.538 ± 0.006	0.810 ± 0.005
Omni-MATH-2-Filtered	Qwen3-4B	0.597 ± 0.012	0.593 ± 0.009	0.762 ± 0.008	0.626 ± 0.017	0.807 ± 0.015	0.870 ± 0.015
Omni-MATH-2-Filtered	Qwen3-8B	0.608 ± 0.023	0.607 ± 0.022	0.712 ± 0.025	0.620 ± 0.012	0.837 ± 0.027	0.875 ± 0.022
Omni-MATH-2-Filtered	Qwen3-14B	0.610 ± 0.015	0.606 ± 0.010	0.671 ± 0.021	0.639 ± 0.012	0.779 ± 0.010	0.876 ± 0.008
Omni-MATH-2-Filtered	gpt-oss-20b	0.563 ± 0.015	0.567 ± 0.016	0.639 ± 0.027	0.604 ± 0.020	0.682 ± 0.019	0.824 ± 0.022
Omni-MATH-2-Filtered	gpt-oss-120b	0.568 ± 0.018	0.568 ± 0.020	0.618 ± 0.013	0.607 ± 0.015	0.671 ± 0.012	0.832 ± 0.019

Appendix Table A33: **Public display-subset AUROC by model–benchmark setting.** Each entry is the held-out test AUROC of the public monitor (rate surface + Cohere) for the corresponding visible display, averaged over five data splits (± SD across data splits). The final row averages the per-setting means across all 15 settings (± between-setting SD).

Benchmark	Model	R	S	C	$R+S$	$R+C$	$S+C$	$R+S+C$
GPQA-Diamond	Qwen3-4B	0.505 ± 0.066	0.735 ± 0.058	0.745 ± 0.065	0.732 ± 0.063	0.738 ± 0.078	0.738 ± 0.062	0.732 ± 0.067
GPQA-Diamond	Qwen3-8B	0.631 ± 0.047	0.803 ± 0.067	0.848 ± 0.058	0.812 ± 0.082	0.846 ± 0.065	0.839 ± 0.065	0.839 ± 0.062
GPQA-Diamond	Qwen3-14B	0.558 ± 0.046	0.746 ± 0.084	0.781 ± 0.068	0.748 ± 0.056	0.791 ± 0.072	0.787 ± 0.046	0.800 ± 0.058
GPQA-Diamond	gpt-oss-20b	0.535 ± 0.137	0.701 ± 0.122	0.725 ± 0.084	0.755 ± 0.066	0.717 ± 0.095	0.708 ± 0.104	0.707 ± 0.096
GPQA-Diamond	gpt-oss-120b	0.542 ± 0.073	0.737 ± 0.062	0.752 ± 0.052	0.761 ± 0.053	0.743 ± 0.062	0.760 ± 0.046	0.770 ± 0.057
MMLU-Pro	Qwen3-4B	0.560 ± 0.014	0.771 ± 0.008	0.824 ± 0.011	0.764 ± 0.009	0.824 ± 0.010	0.824 ± 0.009	0.823 ± 0.010
MMLU-Pro	Qwen3-8B	0.652 ± 0.014	0.780 ± 0.010	0.831 ± 0.005	0.774 ± 0.013	0.819 ± 0.005	0.829 ± 0.005	0.821 ± 0.004
MMLU-Pro	Qwen3-14B	0.533 ± 0.004	0.780 ± 0.008	0.826 ± 0.008	0.775 ± 0.010	0.824 ± 0.008	0.826 ± 0.008	0.825 ± 0.007
MMLU-Pro	gpt-oss-20b	0.515 ± 0.009	0.772 ± 0.015	0.800 ± 0.015	0.767 ± 0.016	0.801 ± 0.014	0.803 ± 0.015	0.803 ± 0.015
MMLU-Pro	gpt-oss-120b	0.538 ± 0.006	0.771 ± 0.004	0.810 ± 0.005	0.769 ± 0.004	0.808 ± 0.005	0.811 ± 0.005	0.810 ± 0.005
Omni-MATH-2-Filtered	Qwen3-4B	0.807 ± 0.015	0.789 ± 0.016	0.866 ± 0.016	0.803 ± 0.018	0.870 ± 0.016	0.866 ± 0.017	0.870 ± 0.015
Omni-MATH-2-Filtered	Qwen3-8B	0.837 ± 0.027	0.838 ± 0.022	0.870 ± 0.019	0.853 ± 0.026	0.865 ± 0.021	0.869 ± 0.022	0.875 ± 0.022
Omni-MATH-2-Filtered	Qwen3-14B	0.779 ± 0.010	0.825 ± 0.003	0.868 ± 0.005	0.821 ± 0.006	0.874 ± 0.008	0.868 ± 0.005	0.876 ± 0.008
Omni-MATH-2-Filtered	gpt-oss-20b	0.682 ± 0.019	0.775 ± 0.016	0.820 ± 0.020	0.775 ± 0.013	0.824 ± 0.020	0.824 ± 0.019	0.824 ± 0.022
Omni-MATH-2-Filtered	gpt-oss-120b	0.671 ± 0.012	0.781 ± 0.023	0.830 ± 0.017	0.776 ± 0.022	0.832 ± 0.018	0.831 ± 0.019	0.832 ± 0.019
Mean		0.623 ± 0.112	0.774 ± 0.035	0.813 ± 0.045	0.779 ± 0.031	0.812 ± 0.048	0.812 ± 0.047	0.814 ± 0.048

Appendix Table A34: **Order-independent Shapley attribution allocates the largest mean AUROC contribution to the trace.** Shapley AUROC contributions of the response R, self-summary S, and reasoning trace C to the full public display, computed over all 3! source orderings with an empty-display baseline of 0.5, averaged over five data splits ($\pm$ SD across data splits). The contributions of a setting sum to its full public AUROC minus 0.5. The final row averages per-setting means across all 15 settings ($\pm$ between-setting SD).

Benchmark	Model	$\phi(R)$	$\phi(S)$	$\phi(C)$	$R{+}S{+}C$ AUROC
GPQA-Diamond	Qwen3-4B	-0.002 ± 0.022	$+0.113 \pm 0.031$	$+0.121 \pm 0.053$	0.732
GPQA-Diamond	Qwen3-8B	$+0.045 \pm 0.018$	$+0.127 \pm 0.030$	$+0.167 \pm 0.020$	0.839
GPQA-Diamond	Qwen3-14B	$+0.026 \pm 0.018$	$+0.118 \pm 0.035$	$+0.157 \pm 0.047$	0.800
GPQA-Diamond	gpt-oss-20b	$+0.019 \pm 0.064$	$+0.097 \pm 0.068$	$+0.090 \pm 0.072$	0.707
GPQA-Diamond	gpt-oss-120b	$+0.020 \pm 0.018$	$+0.126 \pm 0.030$	$+0.125 \pm 0.042$	0.770
MMLU-Pro	Qwen3-4B	$+0.019 \pm 0.004$	$+0.124 \pm 0.005$	$+0.181 \pm 0.007$	0.823
MMLU-Pro	Qwen3-8B	$+0.045 \pm 0.005$	$+0.114 \pm 0.005$	$+0.162 \pm 0.008$	0.821
MMLU-Pro	Qwen3-14B	$+0.010 \pm 0.002$	$+0.134 \pm 0.004$	$+0.182 \pm 0.007$	0.825
MMLU-Pro	gpt-oss-20b	$+0.004 \pm 0.003$	$+0.134 \pm 0.007$	$+0.165 \pm 0.007$	0.803
MMLU-Pro	gpt-oss-120b	$+0.012 \pm 0.002$	$+0.130 \pm 0.003$	$+0.169 \pm 0.004$	0.810
Omni-MATH-2-Filtered	Qwen3-4B	$+0.107 \pm 0.005$	$+0.096 \pm 0.006$	$+0.167 \pm 0.006$	0.870
Omni-MATH-2-Filtered	Qwen3-8B	$+0.116 \pm 0.011$	$+0.119 \pm 0.009$	$+0.140 \pm 0.006$	0.875
Omni-MATH-2-Filtered	Qwen3-14B	$+0.096 \pm 0.006$	$+0.116 \pm 0.002$	$+0.164 \pm 0.006$	0.876
Omni-MATH-2-Filtered	gpt-oss-20b	$+0.061 \pm 0.007$	$+0.108 \pm 0.008$	$+0.155 \pm 0.017$	0.824
Omni-MATH-2-Filtered	gpt-oss-120b	$+0.057 \pm 0.004$	$+0.111 \pm 0.011$	$+0.164 \pm 0.012$	0.832
Mean		$+0.042 \pm 0.038$	$+0.118 \pm 0.012$	$+0.154 \pm 0.025$	0.814 ± 0.048

Appendix Table A35: **Mean AUROC is similar under generic and preservation-tuned summary instructions.** Held-out test AUROC of the public monitor for summary-containing displays when the self-summary is generated with the primary preservation instruction (Pres.) versus a generic instruction requesting compression to at most 90 words (Gen.), averaged over five data splits (± SD across data splits).

Benchmark	Model	S Pres.	S Gen.	$R+S$ Pres.	$R+S$ Gen.	$S+C$ Pres.	$S+C$ Gen.	$R+S+C$ Pres.	$R+S+C$ Gen.
GPQA-Diamond	Qwen3-4B	0.735 ± 0.058	0.693 ± 0.046	0.732 ± 0.063	0.704 ± 0.047	0.738 ± 0.062	0.745 ± 0.060	0.732 ± 0.067	0.741 ± 0.070
GPQA-Diamond	Qwen3-8B	0.803 ± 0.067	0.821 ± 0.072	0.812 ± 0.082	0.810 ± 0.075	0.839 ± 0.065	0.839 ± 0.055	0.839 ± 0.062	0.838 ± 0.059
GPQA-Diamond	Qwen3-14B	0.746 ± 0.084	0.728 ± 0.123	0.748 ± 0.056	0.761 ± 0.109	0.787 ± 0.046	0.787 ± 0.073	0.800 ± 0.058	0.780 ± 0.067
GPQA-Diamond	gpt-oss-20b	0.701 ± 0.122	0.726 ± 0.098	0.755 ± 0.066	0.686 ± 0.105	0.708 ± 0.104	0.718 ± 0.082	0.707 ± 0.096	0.718 ± 0.086
GPQA-Diamond	gpt-oss-120b	0.737 ± 0.062	0.765 ± 0.036	0.761 ± 0.053	0.749 ± 0.038	0.760 ± 0.046	0.767 ± 0.053	0.770 ± 0.057	0.766 ± 0.056
MMLU-Pro	Qwen3-4B	0.771 ± 0.008	0.761 ± 0.013	0.764 ± 0.009	0.758 ± 0.012	0.824 ± 0.009	0.822 ± 0.010	0.823 ± 0.010	0.822 ± 0.011
MMLU-Pro	Qwen3-8B	0.780 ± 0.010	0.778 ± 0.014	0.774 ± 0.013	0.771 ± 0.016	0.829 ± 0.005	0.830 ± 0.006	0.821 ± 0.004	0.822 ± 0.006
MMLU-Pro	Qwen3-14B	0.780 ± 0.008	0.764 ± 0.008	0.775 ± 0.010	0.760 ± 0.008	0.826 ± 0.008	0.824 ± 0.008	0.825 ± 0.007	0.823 ± 0.007
MMLU-Pro	gpt-oss-20b	0.772 ± 0.015	0.767 ± 0.007	0.767 ± 0.016	0.763 ± 0.009	0.803 ± 0.015	0.803 ± 0.013	0.803 ± 0.015	0.802 ± 0.013
MMLU-Pro	gpt-oss-120b	0.771 ± 0.004	0.754 ± 0.007	0.769 ± 0.004	0.751 ± 0.007	0.811 ± 0.005	0.809 ± 0.005	0.810 ± 0.005	0.809 ± 0.003
Omni-MATH-2-Filtered	Qwen3-4B	0.789 ± 0.016	0.783 ± 0.016	0.803 ± 0.018	0.802 ± 0.016	0.866 ± 0.017	0.865 ± 0.016	0.870 ± 0.015	0.869 ± 0.015
Omni-MATH-2-Filtered	Qwen3-8B	0.838 ± 0.022	0.836 ± 0.022	0.853 ± 0.026	0.857 ± 0.028	0.869 ± 0.022	0.866 ± 0.020	0.875 ± 0.022	0.874 ± 0.020
Omni-MATH-2-Filtered	Qwen3-14B	0.825 ± 0.003	0.808 ± 0.010	0.821 ± 0.006	0.810 ± 0.007	0.868 ± 0.005	0.869 ± 0.003	0.876 ± 0.008	0.873 ± 0.007
Omni-MATH-2-Filtered	gpt-oss-20b	0.775 ± 0.016	0.770 ± 0.023	0.775 ± 0.013	0.774 ± 0.021	0.824 ± 0.019	0.824 ± 0.021	0.824 ± 0.022	0.829 ± 0.017
Omni-MATH-2-Filtered	gpt-oss-120b	0.781 ± 0.023	0.771 ± 0.021	0.776 ± 0.022	0.765 ± 0.014	0.831 ± 0.019	0.831 ± 0.019	0.832 ± 0.019	0.832 ± 0.020

Appendix Table A36: **A fixed primary monitor is stable across three sampled summaries and one greedy summary.** For every setting and access level, regularization is selected on the first data split's training and validation items after excluding the balanced 48-item subset, then the surface-plus-Cohere monitor is refit once on the full complement. AUROC columns report the mean across the three sampled summaries and observed settings; brackets are hierarchical 95% bootstrap intervals that resample items within the six correctness-by-trace-length strata and resample observed settings. SD, range, and MAPD summarize the three sampled Δ_S values, where MAPD is the mean absolute pairwise difference. Greedy − sampled compares the greedy Δ_S with the sampled mean. Equal-stratum AUROC is primary; IPW uses the recorded inverse selection probabilities.

Scope	Weighting	S AUROC	$R+S$ AUROC	Δ_S	SD	Range	MAPD	Greedy − sampled	Settings
GPQA-Diamond	Equal-stratum	0.686 [0.596, 0.780]	0.707 [0.622, 0.786]	+0.077 [−0.015, +0.171]	0.017	0.033	0.022	+0.017 [−0.001, +0.038]	5
	IPW	0.779 [0.695, 0.859]	0.797 [0.736, 0.859]	+0.149 [+0.055, +0.248]	0.015	0.027	0.018	+0.015 [−0.001, +0.033]	5
MMLU-Pro	Equal-stratum	0.757 [0.670, 0.834]	0.744 [0.661, 0.827]	+0.153 [+0.052, +0.254]	0.024	0.045	0.030	+0.009 [−0.024, +0.044]	5
	IPW	0.787 [0.711, 0.855]	0.769 [0.695, 0.840]	+0.181 [+0.075, +0.284]	0.031	0.059	0.039	+0.007 [−0.028, +0.042]	5
Omni-MATH-2-Filtered	Equal-stratum	0.653 [0.555, 0.763]	0.675 [0.552, 0.799]	+0.027 [−0.039, +0.089]	0.009	0.017	0.012	+0.003 [−0.014, +0.020]	5
	IPW	0.770 [0.685, 0.848]	0.781 [0.674, 0.881]	+0.051 [−0.015, +0.117]	0.009	0.018	0.012	+0.001 [−0.019, +0.018]	5
All settings	Equal-stratum	0.699 [0.635, 0.756]	0.708 [0.650, 0.765]	+0.086 [+0.027, +0.146]	0.017	0.032	0.021	+0.010 [−0.005, +0.024]	15
	IPW	0.779 [0.732, 0.820]	0.782 [0.731, 0.830]	+0.127 [+0.066, +0.187]	0.018	0.034	0.023	+0.008 [−0.008, +0.023]	15

Appendix Table A37: **The exact 90-word cap and conclusion-region controls use the primary surface-plus-Cohere monitor.** Each AUROC entry is the mean ± between-setting SD of the setting-level held-out test AUROC, where each setting mean averages the five prespecified data splits. Δ is paired within setting against the original uncapped preservation summary. Cap90 deterministically truncates that summary to at most 90 whitespace-delimited words; Last=n_S uses up to the final n_S trace words, where n_S is the capped summary length; Last90 uses up to the final 90 trace words and therefore the whole trace when it is shorter; and Last90 masked removes answer, option, and prompt-number cues before extraction. The final row reports Last90 minus Pres. for the S display; brackets are descriptive 95% setting-bootstrap intervals after averaging the five data splits within each of the 15 observed settings.

Summary slot	S AUROC	S Δ	$R{+}S$ AUROC	$R{+}S$ Δ	n
Pres. (uncapped)	0.774 ± 0.035	0.000 ± 0.000	0.779 ± 0.031	0.000 ± 0.000	15
Cap90	0.775 ± 0.026	+0.001 ± 0.018	0.774 ± 0.034	−0.005 ± 0.007	15
Last=n_S	0.784 ± 0.028	+0.010 ± 0.017	0.779 ± 0.029	0.000 ± 0.012	15
Last90	0.780 ± 0.032	+0.006 ± 0.019	0.778 ± 0.033	−0.001 ± 0.014	15
Last90 masked	0.782 ± 0.031	+0.009 ± 0.022	0.780 ± 0.032	+0.001 ± 0.017	15
Last90 − Pres.	—	+0.006 [−0.003, +0.016]	—	—	15

Appendix Table A38: **Length-matched extractive controls compare summaries with trace extracts at the same word count.** Held-out test AUROC of the open surface-plus-TF-IDF monitor for the summary-only display. Pres. uses the original preservation-tuned self-summary; Cap90 hard-truncates that summary to at most 90 words; First=nS, Last=nS, Head+Tail=nS, and Random=nS use trace words matched to the capped summary length n_S; Last90 uses up to the final 90 trace words and therefore the whole trace when it is shorter; Last90 masked applies answer, option, and prompt-number masking before the same extraction. Values average five data splits (± SD across data splits). The matched mean averages settings where every displayed variant is available. The two contrasts report Last=n_S minus Cap90 and Last90 minus Pres.; brackets are descriptive 95% setting-bootstrap intervals after averaging the five data splits within each of the 15 matched settings.

Benchmark	Model	S Pres.	S Cap90	S First=nS	S Last=nS	S Head+Tail=nS	S Random=nS	S Last90	S Last90 masked
GPQA-Diamond	Qwen3-4B	0.663 ± 0.035	0.659 ± 0.034	0.644 ± 0.066	0.716 ± 0.058	0.646 ± 0.118	0.769 ± 0.092	0.741 ± 0.071	0.720 ± 0.075
GPQA-Diamond	Qwen3-8B	0.750 ± 0.047	0.769 ± 0.050	0.768 ± 0.092	0.833 ± 0.073	0.829 ± 0.054	0.771 ± 0.038	0.830 ± 0.080	0.842 ± 0.058
GPQA-Diamond	Qwen3-14B	0.779 ± 0.081	0.778 ± 0.073	0.720 ± 0.053	0.774 ± 0.094	0.715 ± 0.065	0.820 ± 0.104	0.785 ± 0.088	0.773 ± 0.084
GPQA-Diamond	gpt-oss-20b	0.700 ± 0.080	0.697 ± 0.074	0.722 ± 0.054	0.738 ± 0.114	0.713 ± 0.108	0.762 ± 0.084	0.737 ± 0.117	0.742 ± 0.113
GPQA-Diamond	gpt-oss-120b	0.736 ± 0.100	0.735 ± 0.099	0.718 ± 0.090	0.807 ± 0.036	0.714 ± 0.084	0.760 ± 0.093	0.767 ± 0.047	0.784 ± 0.066
MMLU-Pro	Qwen3-4B	0.802 ± 0.006	0.802 ± 0.006	0.734 ± 0.005	0.814 ± 0.009	0.796 ± 0.005	0.772 ± 0.009	0.824 ± 0.010	0.826 ± 0.010
MMLU-Pro	Qwen3-8B	0.815 ± 0.009	0.803 ± 0.010	0.745 ± 0.008	0.820 ± 0.006	0.815 ± 0.010	0.795 ± 0.009	0.821 ± 0.007	0.823 ± 0.006
MMLU-Pro	Qwen3-14B	0.813 ± 0.006	0.811 ± 0.007	0.745 ± 0.005	0.817 ± 0.007	0.805 ± 0.010	0.777 ± 0.009	0.819 ± 0.004	0.821 ± 0.005
MMLU-Pro	gpt-oss-20b	0.788 ± 0.014	0.784 ± 0.015	0.749 ± 0.014	0.798 ± 0.013	0.790 ± 0.012	0.774 ± 0.010	0.801 ± 0.011	0.805 ± 0.010
MMLU-Pro	gpt-oss-120b	0.802 ± 0.008	0.801 ± 0.009	0.756 ± 0.005	0.807 ± 0.009	0.789 ± 0.004	0.788 ± 0.005	0.812 ± 0.009	0.814 ± 0.009
Omni-MATH-2-Filtered	Qwen3-4B	0.813 ± 0.009	0.799 ± 0.009	0.748 ± 0.013	0.845 ± 0.013	0.839 ± 0.014	0.767 ± 0.011	0.851 ± 0.006	0.838 ± 0.016
Omni-MATH-2-Filtered	Qwen3-8B	0.849 ± 0.020	0.834 ± 0.018	0.806 ± 0.025	0.874 ± 0.014	0.868 ± 0.011	0.810 ± 0.020	0.866 ± 0.014	0.871 ± 0.013
Omni-MATH-2-Filtered	Qwen3-14B	0.845 ± 0.006	0.832 ± 0.006	0.771 ± 0.007	0.853 ± 0.017	0.845 ± 0.015	0.773 ± 0.007	0.840 ± 0.014	0.849 ± 0.019
Omni-MATH-2-Filtered	gpt-oss-20b	0.784 ± 0.005	0.785 ± 0.008	0.727 ± 0.018	0.773 ± 0.016	0.766 ± 0.021	0.755 ± 0.025	0.780 ± 0.012	0.767 ± 0.014
Omni-MATH-2-Filtered	gpt-oss-120b	0.795 ± 0.021	0.789 ± 0.021	0.743 ± 0.021	0.786 ± 0.015	0.789 ± 0.014	0.768 ± 0.016	0.775 ± 0.021	0.772 ± 0.014
Matched mean (15 settings)		0.782	0.778	0.740	0.804	0.781	0.777	0.803	0.803
Last=n_S − Cap90									+0.025 [+0.013, +0.038]
Last90 − Pres.									+0.021 [+0.008, +0.036]

Appendix Table A39: **Chunk-based trace monitors are compared directly with the matched summary-only and trace-only monitors.** AUROC values are means ± between-setting SD after first averaging the five prespecified data splits within each of the 15 model–benchmark settings. Deltas are paired by setting against the primary summary-only (S) and trace-only (C) monitors. Each full trace is divided into balanced contiguous word chunks targeting 512 words and capped at eight chunks. The Cohere variants pool fixed 512-dimensional chunk vectors or fit a linear-attention multiple-instance head. The corresponding surface-plus-TF–IDF analysis is reported in Appendix Table A40.

Monitor	AUROC	Δ vs S	Δ vs C	Settings
Primary whole-display S	0.774 ± 0.035	+0.000 ± 0.000	−0.039 ± 0.017	15
Primary whole-display C	0.813 ± 0.045	+0.039 ± 0.017	+0.000 ± 0.000	15
Cohere chunks: mean pooling	0.812 ± 0.048	+0.038 ± 0.020	−0.001 ± 0.008	15
Cohere chunks: elementwise max	0.811 ± 0.044	+0.038 ± 0.019	−0.002 ± 0.010	15
Cohere chunks: elementwise top-2	0.813 ± 0.043	+0.040 ± 0.018	+0.000 ± 0.011	15
Cohere chunks: attention MIL	0.792 ± 0.049	+0.018 ± 0.021	−0.021 ± 0.013	15

Appendix Table A40: **Scores from trace segments are compared with scores from complete summaries and traces on the full dataset.** AUROC values first average the five prespecified random splits within each model–benchmark combination and then report the mean ± SD across all 15 combinations. Deltas compare values within the same combination with the complete summary and trace, using the same surface features and TF–IDF features. The segment and complete-text classifiers use the same six candidate inverse-regularization values, $C \in \{0.0001, 0.001, 0.01, 0.1, 1, 10\}$. Each trace is divided into up to eight contiguous segments of similar length, targeting 512 words. The analysis covers every item assigned to each data split. Within each split, C is selected by validation item-level AUROC, and scores are combined only for held-out test items. The top-two mean, mean, and last-segment results use that same value of C. These alternatives test sensitivity to how traces are divided and segment scores are combined; they do not locate causal evidence. A short trace may form one segment; during fitting, each item is weighted once for every segment it contains; and maximum or top-two aggregation gives longer traces more opportunities to receive a high score.

Input and score	AUROC	Δ vs S	Δ vs C	Settings
Whole-display summary S	0.782 ± 0.051	+0.000 ± 0.000	−0.036 ± 0.021	15
Whole-display trace C	0.818 ± 0.051	+0.036 ± 0.021	+0.000 ± 0.000	15
Segment scores: maximum	0.799 ± 0.051	+0.017 ± 0.029	−0.019 ± 0.016	15
Segment scores: top-two mean	0.803 ± 0.051	+0.021 ± 0.024	−0.015 ± 0.013	15
Segment scores: mean	0.818 ± 0.046	+0.036 ± 0.026	+0.000 ± 0.012	15
Last (or only) segment	0.817 ± 0.043	+0.035 ± 0.026	−0.000 ± 0.015	15

Appendix Table A41: **Public-text feature comparisons.** AUROC values are mean ± descriptive SD across the model–benchmark settings and data splits available at each public access level. These SDs are not confidence intervals. All displayed feature families cover the 15 model–benchmark settings. The final column gives the number of model–benchmark–split estimates.

Access level	Feature family	AUROC	Estimates
R	Surface	0.584 ± 0.098	75
R	No length	0.579 ± 0.092	75
R	Cohere	0.620 ± 0.113	75
R	Surface + Cohere	0.623 ± 0.117	75
R	TF-IDF	0.618 ± 0.121	75
R	Surface + TF-IDF + Cohere	0.623 ± 0.117	75
$R + S$	Surface	0.722 ± 0.054	75
$R + S$	No length	0.701 ± 0.052	75
$R + S$	Cohere	0.773 ± 0.046	75
$R + S$	Surface + Cohere	0.779 ± 0.047	75
$R + S$	TF-IDF	0.797 ± 0.050	75
$R + S$	Surface + TF-IDF + Cohere	0.784 ± 0.047	75
$R + S + C$	Surface	0.803 ± 0.050	75
$R + S + C$	No length	0.777 ± 0.043	75
$R + S + C$	Cohere	0.802 ± 0.059	75
$R + S + C$	Surface + Cohere	0.814 ± 0.060	75
$R + S + C$	TF-IDF	0.815 ± 0.048	75
$R + S + C$	Surface + TF-IDF + Cohere	0.814 ± 0.060	75

Appendix Table A42: **Correctness ranking remains after adjusting the other features for text length.** For every setting, access level, and data split, word and sentence counts are nuisance covariates. Inside each fitting fold, multi-output least squares residualizes every other primary surface and Cohere feature on those two counts; the counts are then excluded before standardization and logistic-probe fitting. Values average the five data splits within each setting and then give equal weight to the 15 settings. Change is the adjusted estimate minus the original primary-monitor estimate.

Quantity	Primary monitor	After length adjustment	Change
Access-level AUROC			
R	0.623	0.617	−0.006
S	0.774	0.689	−0.084
C	0.813	0.716	−0.097
$R{+}S$	0.779	0.734	−0.045
$R{+}S{+}C$	0.814	0.717	−0.097
$q{+}R$	0.750	0.727	−0.023
$q{+}R{+}S$	0.769	0.736	−0.033
$q{+}R{+}C$	0.812	0.712	−0.100
$q{+}R{+}S{+}C$	0.809	0.713	−0.097
Paired AUROC increments			
Δ_S	0.156	0.117	−0.039
Δ_C	0.035	−0.017	−0.052
Δ_{public}	0.191	0.101	−0.090
$q\Delta_S$	0.019	0.010	−0.009
$q\Delta_{C\|R}$	0.062	−0.015	−0.077
$q\Delta_C$	0.041	−0.023	−0.064
$q\Delta_{S\|R,C}$	−0.002	0.001	+0.003
$q(C{-}S)$	0.043	−0.024	−0.067
$q\Delta_{\text{public}}$	0.060	−0.014	−0.073

Appendix Table A43: **Public embeddings fit without silent truncation.** Cohere embed-v4.0 was applied with silent truncation disabled. Counts refer to items at the full public level $R + S + C$, pooled across the five target models for each benchmark.

Benchmark	$R + S + C$ items	Whole-context fit	Mean chunks	Max chunks
GPQA-Diamond	990	100%	1.00	1
MMLU-Pro	60,160	100%	1.00	1
Omni-MATH-2-Filtered	20,905	100%	1.00	1

Appendix Table A44: **Stripping prompt-overlapping text from traces and summaries leaves public AUROC essentially unchanged.** Paired AUROC change when 8–14-word spans copied from the task prompt are removed from the visible summary and trace before feature extraction, averaged over five data splits (± SD across data splits). The final row averages per-setting means across all 15 settings (± between-setting SD).

Benchmark	Model	S	C	$R{+}S$	$R{+}C$	$S{+}C$	$R{+}S{+}C$
GPQA-Diamond	Qwen3-4B	+0.006 ± 0.020	−0.004 ± 0.019	+0.016 ± 0.014	−0.008 ± 0.019	−0.005 ± 0.012	−0.003 ± 0.021
GPQA-Diamond	Qwen3-8B	−0.003 ± 0.010	+0.007 ± 0.012	−0.006 ± 0.010	+0.004 ± 0.005	+0.003 ± 0.005	+0.000 ± 0.009
GPQA-Diamond	Qwen3-14B	+0.019 ± 0.046	+0.014 ± 0.030	−0.013 ± 0.019	−0.019 ± 0.033	−0.014 ± 0.053	−0.021 ± 0.037
GPQA-Diamond	gpt-oss-20b	−0.027 ± 0.067	−0.004 ± 0.029	−0.030 ± 0.080	+0.007 ± 0.042	+0.002 ± 0.014	−0.001 ± 0.020
GPQA-Diamond	gpt-oss-120b	+0.005 ± 0.016	−0.002 ± 0.020	+0.004 ± 0.007	−0.002 ± 0.032	+0.005 ± 0.020	−0.007 ± 0.019
MMLU-Pro	Qwen3-4B	+0.000 ± 0.001	+0.000 ± 0.003	+0.000 ± 0.000	+0.000 ± 0.003	−0.000 ± 0.002	−0.000 ± 0.002
MMLU-Pro	Qwen3-8B	−0.000 ± 0.001	+0.001 ± 0.001	+0.000 ± 0.000	+0.001 ± 0.001	−0.000 ± 0.001	−0.000 ± 0.001
MMLU-Pro	Qwen3-14B	+0.000 ± 0.001	+0.000 ± 0.001	+0.001 ± 0.001	−0.000 ± 0.001	+0.000 ± 0.001	−0.001 ± 0.001
MMLU-Pro	gpt-oss-20b	+0.002 ± 0.000	−0.000 ± 0.003	+0.002 ± 0.000	−0.001 ± 0.002	−0.001 ± 0.001	−0.001 ± 0.001
MMLU-Pro	gpt-oss-120b	−0.000 ± 0.000	+0.000 ± 0.002	−0.001 ± 0.001	−0.001 ± 0.002	+0.001 ± 0.003	+0.000 ± 0.003
Omni-MATH-2-Filtered	Qwen3-4B	−0.001 ± 0.003	−0.001 ± 0.002	+0.004 ± 0.002	+0.001 ± 0.001	+0.002 ± 0.003	+0.000 ± 0.002
Omni-MATH-2-Filtered	Qwen3-8B	+0.001 ± 0.002	−0.001 ± 0.002	+0.001 ± 0.001	−0.003 ± 0.005	+0.003 ± 0.002	+0.000 ± 0.002
Omni-MATH-2-Filtered	Qwen3-14B	−0.001 ± 0.002	+0.003 ± 0.003	+0.001 ± 0.001	−0.001 ± 0.001	+0.003 ± 0.001	+0.001 ± 0.002
Omni-MATH-2-Filtered	gpt-oss-20b	−0.005 ± 0.008	−0.001 ± 0.007	−0.001 ± 0.001	+0.001 ± 0.002	−0.001 ± 0.007	−0.000 ± 0.002
Omni-MATH-2-Filtered	gpt-oss-120b	−0.003 ± 0.007	+0.000 ± 0.003	+0.001 ± 0.001	−0.002 ± 0.001	+0.000 ± 0.002	+0.001 ± 0.001
Mean		−0.000 ± 0.009	+0.001 ± 0.005	−0.001 ± 0.010	−0.001 ± 0.006	−0.000 ± 0.005	−0.002 ± 0.006

Appendix Table A45: **Masking of answers, options, and prompt-derived numbers leaves public-text AUROC essentially unchanged.** Paired AUROC change relative to the unmasked display when answer spans and choice letters, option texts, and prompt-derived numeric constants are replaced by fixed markers before feature extraction (`maskall` applies all three), averaged over five data splits (± SD across data splits). The three rightmost columns ablate one component at a time on the trace display. Masking bounds the direct and near-direct reconstruction channel, but it does not remove paraphrase or item-level cues and can introduce marker artefacts. The final row averages per-setting means across settings (± between-setting SD).

Benchmark	Model	S maskall	C maskall	R+S+C maskall	C (options)	C (answer)	C (numbers)
GPQA-Diamond	Qwen3-4B	−0.012 ± 0.027	−0.008 ± 0.008	−0.002 ± 0.013	−0.013 ± 0.022	+0.006 ± 0.011	−0.012 ± 0.009
GPQA-Diamond	Qwen3-8B	+0.011 ± 0.025	+0.010 ± 0.008	−0.002 ± 0.008	+0.007 ± 0.006	+0.004 ± 0.005	+0.004 ± 0.004
GPQA-Diamond	Qwen3-14B	+0.033 ± 0.030	+0.004 ± 0.011	+0.010 ± 0.015	−0.002 ± 0.034	−0.007 ± 0.032	−0.001 ± 0.026
GPQA-Diamond	gpt-oss-20b	+0.009 ± 0.008	−0.007 ± 0.039	+0.002 ± 0.007	−0.012 ± 0.017	−0.003 ± 0.029	−0.026 ± 0.044
GPQA-Diamond	gpt-oss-120b	+0.020 ± 0.028	+0.001 ± 0.010	+0.003 ± 0.005	+0.002 ± 0.009	+0.004 ± 0.009	+0.010 ± 0.006
MMLU-Pro	Qwen3-4B	+0.001 ± 0.002	+0.002 ± 0.001	+0.002 ± 0.001	−0.000 ± 0.001	+0.001 ± 0.001	+0.001 ± 0.001
MMLU-Pro	Qwen3-8B	+0.000 ± 0.002	+0.002 ± 0.001	+0.003 ± 0.001	+0.000 ± 0.001	+0.001 ± 0.001	+0.000 ± 0.001
MMLU-Pro	Qwen3-14B	−0.000 ± 0.002	+0.000 ± 0.002	+0.000 ± 0.003	−0.000 ± 0.001	+0.000 ± 0.001	−0.000 ± 0.001
MMLU-Pro	gpt-oss-20b	+0.003 ± 0.002	+0.002 ± 0.002	+0.003 ± 0.001	+0.000 ± 0.001	+0.001 ± 0.001	+0.001 ± 0.001
MMLU-Pro	gpt-oss-120b	+0.003 ± 0.003	−0.001 ± 0.004	+0.002 ± 0.002	+0.000 ± 0.001	−0.001 ± 0.001	+0.000 ± 0.003
Omni-MATH-2-Filtered	Qwen3-4B	+0.002 ± 0.005	+0.000 ± 0.002	−0.002 ± 0.002	+0.000 ± 0.000	−0.001 ± 0.001	−0.001 ± 0.001
Omni-MATH-2-Filtered	Qwen3-8B	−0.001 ± 0.002	−0.001 ± 0.003	−0.001 ± 0.001	+0.000 ± 0.000	−0.002 ± 0.002	+0.000 ± 0.003
Omni-MATH-2-Filtered	Qwen3-14B	+0.003 ± 0.003	+0.002 ± 0.003	+0.001 ± 0.003	+0.000 ± 0.000	−0.000 ± 0.001	+0.002 ± 0.003
Omni-MATH-2-Filtered	gpt-oss-20b	+0.002 ± 0.004	+0.003 ± 0.001	+0.001 ± 0.002	+0.000 ± 0.000	+0.000 ± 0.001	+0.002 ± 0.001
Omni-MATH-2-Filtered	gpt-oss-120b	−0.001 ± 0.003	+0.000 ± 0.002	−0.004 ± 0.007	−0.000 ± 0.000	−0.001 ± 0.001	+0.001 ± 0.002
Mean		+0.005 ± 0.010	+0.001 ± 0.004	+0.001 ± 0.003	−0.001 ± 0.005	+0.000 ± 0.003	−0.001 ± 0.008

Appendix Table A46: **Adding the task prompt does not improve mean full-display AUROC.** Held-out test AUROC of the public monitor for prompt-conditioned control displays, averaged over five data splits (± SD across data splits), with the output-only $R + S + C$ display repeated for reference. The final row averages per-setting means across all 15 settings (± between-setting SD).

Benchmark	Model	$R+S+C$	q	$q+R$	$q+S$	$q+C$	$q+R+S$	$q+R+C$	$q+S+C$	$q+R+S+C$
GPQA-Diamond	Qwen3-4B	0.732 ± 0.067	0.665 ± 0.077	0.685 ± 0.083	0.701 ± 0.080	0.727 ± 0.079	0.706 ± 0.066	0.731 ± 0.084	0.724 ± 0.078	0.725 ± 0.084
GPQA-Diamond	Qwen3-8B	0.839 ± 0.062	0.815 ± 0.070	0.812 ± 0.067	0.818 ± 0.069	0.840 ± 0.058	0.819 ± 0.071	0.849 ± 0.061	0.839 ± 0.056	0.836 ± 0.058
GPQA-Diamond	Qwen3-14B	0.800 ± 0.058	0.749 ± 0.056	0.750 ± 0.081	0.746 ± 0.068	0.767 ± 0.064	0.742 ± 0.069	0.772 ± 0.067	0.763 ± 0.058	0.758 ± 0.056
GPQA-Diamond	gpt-oss-20b	0.707 ± 0.096	0.753 ± 0.076	0.747 ± 0.087	0.753 ± 0.079	0.759 ± 0.080	0.745 ± 0.081	0.749 ± 0.086	0.721 ± 0.084	0.723 ± 0.077
GPQA-Diamond	gpt-oss-120b	0.770 ± 0.057	0.740 ± 0.066	0.727 ± 0.068	0.717 ± 0.059	0.742 ± 0.053	0.722 ± 0.063	0.740 ± 0.063	0.742 ± 0.048	0.747 ± 0.059
MMLU-Pro	Qwen3-4B	0.823 ± 0.010	0.730 ± 0.013	0.733 ± 0.018	0.758 ± 0.014	0.824 ± 0.009	0.761 ± 0.010	0.824 ± 0.010	0.825 ± 0.009	0.824 ± 0.009
MMLU-Pro	Qwen3-8B	0.821 ± 0.004	0.714 ± 0.014	0.736 ± 0.014	0.768 ± 0.013	0.821 ± 0.006	0.773 ± 0.011	0.820 ± 0.005	0.826 ± 0.004	0.822 ± 0.004
MMLU-Pro	Qwen3-14B	0.825 ± 0.007	0.735 ± 0.011	0.731 ± 0.010	0.772 ± 0.006	0.825 ± 0.007	0.768 ± 0.008	0.825 ± 0.007	0.826 ± 0.006	0.826 ± 0.006
MMLU-Pro	gpt-oss-20b	0.803 ± 0.015	0.710 ± 0.011	0.712 ± 0.009	0.748 ± 0.014	0.802 ± 0.014	0.748 ± 0.014	0.802 ± 0.014	0.804 ± 0.015	0.804 ± 0.015
MMLU-Pro	gpt-oss-120b	0.810 ± 0.005	0.702 ± 0.007	0.702 ± 0.005	0.749 ± 0.002	0.807 ± 0.005	0.747 ± 0.004	0.806 ± 0.005	0.810 ± 0.003	0.809 ± 0.004
Omni-MATH-2-Filtered	Qwen3-4B	0.870 ± 0.015	0.751 ± 0.021	0.805 ± 0.017	0.790 ± 0.019	0.868 ± 0.015	0.803 ± 0.020	0.871 ± 0.015	0.868 ± 0.014	0.870 ± 0.015
Omni-MATH-2-Filtered	Qwen3-8B	0.875 ± 0.022	0.746 ± 0.024	0.825 ± 0.028	0.839 ± 0.023	0.874 ± 0.016	0.844 ± 0.029	0.866 ± 0.019	0.877 ± 0.021	0.870 ± 0.017
Omni-MATH-2-Filtered	Qwen3-14B	0.876 ± 0.008	0.721 ± 0.007	0.796 ± 0.008	0.817 ± 0.006	0.865 ± 0.006	0.813 ± 0.008	0.864 ± 0.007	0.870 ± 0.004	0.871 ± 0.006
Omni-MATH-2-Filtered	gpt-oss-20b	0.824 ± 0.022	0.735 ± 0.016	0.740 ± 0.017	0.766 ± 0.012	0.822 ± 0.018	0.764 ± 0.015	0.825 ± 0.014	0.823 ± 0.017	0.824 ± 0.016
Omni-MATH-2-Filtered	gpt-oss-120b	0.832 ± 0.019	0.742 ± 0.014	0.743 ± 0.006	0.779 ± 0.019	0.832 ± 0.015	0.775 ± 0.016	0.832 ± 0.015	0.830 ± 0.018	0.832 ± 0.016
Mean		0.814 ± 0.048	0.734 ± 0.032	0.750 ± 0.042	0.768 ± 0.037	0.812 ± 0.045	0.769 ± 0.038	0.812 ± 0.045	0.810 ± 0.050	0.809 ± 0.050

Appendix Table A47: **Displays that include the trace add signal across the natural mix of questions, beyond the prompt-only baseline.** Paired held-out AUROC contrasts of prompt-conditioned public displays against the prompt-only and output-only baselines, averaged over five data splits (± SD across data splits). The final row averages per-setting means across all 15 settings (± between-setting SD).

Benchmark	Model	$q+C-q$	$q+R+S+C-q$	$q+C-C$	$q+R+S+C-R+S+C$
GPQA-Diamond	Qwen3-4B	+0.062 ± 0.014	+0.060 ± 0.015	−0.018 ± 0.019	−0.008 ± 0.025
GPQA-Diamond	Qwen3-8B	+0.025 ± 0.028	+0.021 ± 0.034	−0.008 ± 0.009	−0.003 ± 0.011
GPQA-Diamond	Qwen3-14B	+0.019 ± 0.029	+0.009 ± 0.018	−0.013 ± 0.035	−0.042 ± 0.027
GPQA-Diamond	gpt-oss-20b	+0.007 ± 0.034	−0.029 ± 0.044	+0.035 ± 0.069	+0.017 ± 0.028
GPQA-Diamond	gpt-oss-120b	+0.002 ± 0.017	+0.007 ± 0.024	−0.010 ± 0.025	−0.023 ± 0.018
MMLU-Pro	Qwen3-4B	+0.094 ± 0.008	+0.094 ± 0.008	−0.000 ± 0.002	+0.000 ± 0.002
MMLU-Pro	Qwen3-8B	+0.107 ± 0.015	+0.108 ± 0.014	−0.010 ± 0.003	+0.001 ± 0.001
MMLU-Pro	Qwen3-14B	+0.090 ± 0.012	+0.091 ± 0.012	−0.001 ± 0.003	+0.000 ± 0.002
MMLU-Pro	gpt-oss-20b	+0.092 ± 0.008	+0.094 ± 0.009	+0.003 ± 0.001	+0.001 ± 0.002
MMLU-Pro	gpt-oss-120b	+0.105 ± 0.008	+0.107 ± 0.007	−0.002 ± 0.002	−0.001 ± 0.003
Omni-MATH-2-Filtered	Qwen3-4B	+0.116 ± 0.016	+0.119 ± 0.015	+0.002 ± 0.003	+0.001 ± 0.001
Omni-MATH-2-Filtered	Qwen3-8B	+0.128 ± 0.014	+0.124 ± 0.015	+0.003 ± 0.005	−0.006 ± 0.006
Omni-MATH-2-Filtered	Qwen3-14B	+0.143 ± 0.009	+0.150 ± 0.008	−0.004 ± 0.006	−0.005 ± 0.003
Omni-MATH-2-Filtered	gpt-oss-20b	+0.087 ± 0.022	+0.089 ± 0.021	+0.002 ± 0.008	−0.000 ± 0.012
Omni-MATH-2-Filtered	gpt-oss-120b	+0.090 ± 0.016	+0.090 ± 0.017	+0.002 ± 0.004	−0.000 ± 0.003
Mean		+0.078 ± 0.045	+0.076 ± 0.051	−0.001 ± 0.012	−0.004 ± 0.013

Appendix Table A48: **Prompt-conditioned Shapley attribution averages each source over every addition order.** Shapley AUROC contributions of the response R, self-summary S, and reasoning trace C conditional on the prompt q, computed from the complete 2^3 prompt-conditioned factorial over all 3! source orderings. Values average five data splits ($\pm$ SD across data splits). The contributions of a setting sum to the $q{+}R{+}S{+}C$ AUROC minus its q-only AUROC. The final row averages per-setting means across all 15 settings ($\pm$ between-setting SD).

Benchmark	Model	$\phi(R)$	$\phi(S)$	$\phi(C)$	$q{+}R{+}S{+}C$ AUROC
GPQA-Diamond	Qwen3-4B	+0.008 ± 0.010	+0.013 ± 0.006	+0.039 ± 0.012	0.725
GPQA-Diamond	Qwen3-8B	0.000 ± 0.007	−0.002 ± 0.009	+0.024 ± 0.026	0.836
GPQA-Diamond	Qwen3-14B	−0.001 ± 0.011	−0.008 ± 0.025	+0.018 ± 0.032	0.758
GPQA-Diamond	gpt-oss-20b	−0.004 ± 0.006	−0.015 ± 0.033	−0.010 ± 0.021	0.723
GPQA-Diamond	gpt-oss-120b	−0.002 ± 0.011	−0.006 ± 0.012	+0.015 ± 0.011	0.747
MMLU-Pro	Qwen3-4B	+0.001 ± 0.003	+0.014 ± 0.003	+0.079 ± 0.006	0.824
MMLU-Pro	Qwen3-8B	+0.007 ± 0.002	+0.026 ± 0.004	+0.076 ± 0.014	0.822
MMLU-Pro	Qwen3-14B	−0.002 ± 0.001	+0.019 ± 0.003	+0.074 ± 0.010	0.826
MMLU-Pro	gpt-oss-20b	+0.001 ± 0.002	+0.019 ± 0.004	+0.074 ± 0.007	0.804
MMLU-Pro	gpt-oss-120b	−0.001 ± 0.001	+0.025 ± 0.003	+0.083 ± 0.006	0.809
Omni-MATH-2-Filtered	Qwen3-4B	+0.021 ± 0.006	+0.012 ± 0.002	+0.085 ± 0.009	0.870
Omni-MATH-2-Filtered	Qwen3-8B	+0.024 ± 0.005	+0.036 ± 0.006	+0.064 ± 0.010	0.870
Omni-MATH-2-Filtered	Qwen3-14B	+0.025 ± 0.004	+0.038 ± 0.003	+0.087 ± 0.006	0.871
Omni-MATH-2-Filtered	gpt-oss-20b	+0.002 ± 0.003	+0.014 ± 0.007	+0.072 ± 0.017	0.824
Omni-MATH-2-Filtered	gpt-oss-120b	+0.001 ± 0.004	+0.017 ± 0.005	+0.072 ± 0.015	0.832
Mean		+0.005 ± 0.010	+0.013 ± 0.015	+0.057 ± 0.031	0.809 ± 0.050

Appendix Table A49: **An open prompt-conditioned factorial covers every subset of response, summary, and trace.** Held-out AUROC for the surface-feature-plus-TF-IDF monitor, conditional on the task prompt q. Values average five data splits (± SD across data splits). The final row averages the 15 setting means (± between-setting SD). This is an open-encoder control, not the primary surface-plus-Cohere monitor.

Benchmark	Model	q	$q+R$	$q+S$	$q+C$	$q+R+S$	$q+R+C$	$q+S+C$	$q+R+S+C$
GPQA-Diamond	Qwen3-4B	0.646 ± 0.048	0.663 ± 0.061	0.667 ± 0.028	0.721 ± 0.057	0.679 ± 0.041	0.722 ± 0.057	0.724 ± 0.060	0.724 ± 0.060
GPQA-Diamond	Qwen3-8B	0.754 ± 0.069	0.763 ± 0.047	0.773 ± 0.061	0.825 ± 0.025	0.720 ± 0.055	0.827 ± 0.032	0.825 ± 0.027	0.824 ± 0.032
GPQA-Diamond	Qwen3-14B	0.654 ± 0.111	0.649 ± 0.118	0.737 ± 0.116	0.796 ± 0.048	0.735 ± 0.120	0.797 ± 0.048	0.798 ± 0.053	0.799 ± 0.054
GPQA-Diamond	gpt-oss-20b	0.671 ± 0.132	0.668 ± 0.134	0.717 ± 0.080	0.722 ± 0.078	0.716 ± 0.081	0.721 ± 0.078	0.723 ± 0.073	0.724 ± 0.073
GPQA-Diamond	gpt-oss-120b	0.662 ± 0.064	0.662 ± 0.064	0.736 ± 0.080	0.788 ± 0.035	0.735 ± 0.080	0.790 ± 0.034	0.786 ± 0.031	0.786 ± 0.033
MMLU-Pro	Qwen3-4B	0.711 ± 0.020	0.718 ± 0.022	0.789 ± 0.005	0.841 ± 0.008	0.790 ± 0.007	0.841 ± 0.008	0.841 ± 0.008	0.841 ± 0.008
MMLU-Pro	Qwen3-8B	0.702 ± 0.014	0.741 ± 0.014	0.795 ± 0.009	0.843 ± 0.008	0.799 ± 0.011	0.844 ± 0.007	0.845 ± 0.007	0.844 ± 0.007
MMLU-Pro	Qwen3-14B	0.710 ± 0.008	0.712 ± 0.007	0.793 ± 0.005	0.837 ± 0.006	0.794 ± 0.004	0.837 ± 0.006	0.837 ± 0.007	0.838 ± 0.007
MMLU-Pro	gpt-oss-20b	0.693 ± 0.012	0.692 ± 0.015	0.768 ± 0.015	0.806 ± 0.013	0.768 ± 0.015	0.806 ± 0.013	0.808 ± 0.012	0.808 ± 0.012
MMLU-Pro	gpt-oss-120b	0.690 ± 0.007	0.690 ± 0.007	0.780 ± 0.005	0.811 ± 0.003	0.781 ± 0.005	0.811 ± 0.003	0.813 ± 0.003	0.813 ± 0.003
Omni-MATH-2-Filtered	Qwen3-4B	0.740 ± 0.011	0.806 ± 0.014	0.809 ± 0.008	0.877 ± 0.013	0.815 ± 0.014	0.876 ± 0.012	0.876 ± 0.013	0.876 ± 0.012
Omni-MATH-2-Filtered	Qwen3-8B	0.745 ± 0.020	0.843 ± 0.015	0.849 ± 0.015	0.885 ± 0.010	0.844 ± 0.019	0.884 ± 0.011	0.884 ± 0.011	0.883 ± 0.011
Omni-MATH-2-Filtered	Qwen3-14B	0.707 ± 0.017	0.794 ± 0.012	0.835 ± 0.010	0.871 ± 0.005	0.822 ± 0.008	0.868 ± 0.005	0.872 ± 0.006	0.872 ± 0.004
Omni-MATH-2-Filtered	gpt-oss-20b	0.704 ± 0.015	0.727 ± 0.024	0.776 ± 0.006	0.818 ± 0.021	0.784 ± 0.008	0.818 ± 0.020	0.819 ± 0.021	0.818 ± 0.020
Omni-MATH-2-Filtered	gpt-oss-120b	0.718 ± 0.011	0.738 ± 0.015	0.790 ± 0.018	0.833 ± 0.019	0.784 ± 0.018	0.834 ± 0.020	0.834 ± 0.019	0.834 ± 0.020
Mean		0.700 ± 0.032	0.724 ± 0.058	0.774 ± 0.046	0.818 ± 0.048	0.771 ± 0.045	0.818 ± 0.048	0.819 ± 0.048	0.819 ± 0.047

Appendix Table A50: **The complete open factorial shows context-dependent marginal gains.** Mean paired AUROC changes across the 15 settings for the surface-feature-plus-TF-IDF monitor. Intervals are descriptive 95% setting-bootstrap intervals over the observed settings, not population-generalization intervals; the last column counts settings with a positive mean contrast. Summary-first and trace-first marginal gains use the same held-out items and data splits. Here, 'after' denotes set inclusion in the factorial contrast, not a change in the physical presentation order of the displayed text.

Contrast	Mean Δ AUROC	95% descriptive interval	Positive settings
Add S after $q+R$	+0.047	[+0.027, +0.064]	14/15
Add C after $q+R+S$	+0.048	[+0.039, +0.059]	15/15
Add C after $q+R$	+0.094	[+0.078, +0.110]	15/15
Add S after $q+R+C$	+0.000	[−0.001, +0.001]	12/15
$q+R+C$ minus $q+R+S$	+0.047	[+0.038, +0.058]	15/15
$q+R+S+C$ minus $q+R$	+0.094	[+0.079, +0.110]	15/15

Appendix Table A51: **Open prompt-conditioned Shapley values average marginal gains over every inclusion order.** Contributions of response R, summary S, and trace C are computed from the complete surface-feature-plus-TF-IDF factorial, with q as the baseline. Values average five data splits (± SD across data splits); the final row averages the 15 setting means (± between-setting SD). Contributions sum to the public gain over q but remain an attribution over fitted probes rather than an information decomposition. Inclusion order is an attribution path over component sets, not presentation order.

Benchmark	Model	$\phi(R)$	$\phi(S)$	$\phi(C)$	Gain over q	Full AUROC
GPQA-Diamond	Qwen3-4B	+0.008 ± 0.009	+0.011 ± 0.010	+0.059 ± 0.024	+0.078 ± 0.032	0.724 ± 0.060
GPQA-Diamond	Qwen3-8B	−0.006 ± 0.032	−0.002 ± 0.015	+0.077 ± 0.052	+0.069 ± 0.067	0.824 ± 0.032
GPQA-Diamond	Qwen3-14B	−0.002 ± 0.004	+0.043 ± 0.019	+0.104 ± 0.097	+0.145 ± 0.099	0.799 ± 0.054
GPQA-Diamond	gpt-oss-20b	−0.001 ± 0.003	+0.025 ± 0.055	+0.030 ± 0.068	+0.053 ± 0.094	0.724 ± 0.073
GPQA-Diamond	gpt-oss-120b	+0.000 ± 0.001	+0.035 ± 0.023	+0.089 ± 0.046	+0.124 ± 0.051	0.786 ± 0.033
MMLU-Pro	Qwen3-4B	+0.003 ± 0.001	+0.038 ± 0.008	+0.089 ± 0.011	+0.130 ± 0.017	0.841 ± 0.008
MMLU-Pro	Qwen3-8B	+0.013 ± 0.004	+0.041 ± 0.004	+0.087 ± 0.014	+0.142 ± 0.018	0.844 ± 0.007
MMLU-Pro	Qwen3-14B	+0.001 ± 0.000	+0.042 ± 0.004	+0.085 ± 0.010	+0.128 ± 0.013	0.838 ± 0.007
MMLU-Pro	gpt-oss-20b	−0.001 ± 0.002	+0.039 ± 0.003	+0.077 ± 0.005	+0.115 ± 0.005	0.808 ± 0.012
MMLU-Pro	gpt-oss-120b	+0.000 ± 0.000	+0.046 ± 0.002	+0.077 ± 0.005	+0.123 ± 0.007	0.813 ± 0.003
Omni-MATH-2-Filtered	Qwen3-4B	+0.023 ± 0.005	+0.024 ± 0.002	+0.089 ± 0.007	+0.136 ± 0.010	0.876 ± 0.012
Omni-MATH-2-Filtered	Qwen3-8B	+0.031 ± 0.003	+0.035 ± 0.004	+0.073 ± 0.012	+0.139 ± 0.016	0.883 ± 0.011
Omni-MATH-2-Filtered	Qwen3-14B	+0.027 ± 0.003	+0.048 ± 0.005	+0.089 ± 0.011	+0.164 ± 0.017	0.872 ± 0.004
Omni-MATH-2-Filtered	gpt-oss-20b	+0.009 ± 0.008	+0.034 ± 0.008	+0.072 ± 0.024	+0.114 ± 0.030	0.818 ± 0.020
Omni-MATH-2-Filtered	gpt-oss-120b	+0.006 ± 0.006	+0.032 ± 0.008	+0.079 ± 0.017	+0.117 ± 0.023	0.834 ± 0.020
Mean		+0.007 ± 0.011	+0.033 ± 0.014	+0.078 ± 0.017	+0.118 ± 0.030	0.819 ± 0.047

Appendix Table A52: **Under the primary surface-plus-Cohere monitor, summaries remain near chance while traces retain modest discrimination.** Entries give across-item AUROC / within-item AUROC, with 95% item-level bootstrap intervals in square brackets for the within-item value. Conditional-logit entries give the coefficient for a one-SD increase in monitor score and its 95% interval. Both AUROCs use the same post-outcome discordant held-out MMLU-Pro items across three independent runs per item. Monitors are trained on the primary run. Across-item AUROC pools runs across items; within-item AUROC compares a correct and an incorrect run of the same item and model. Pairing holds item-level cues fixed, and the prompt is not supplied to the monitor. There are 939–1,389 discordant items per model.

Prompt-withheld displays							
Model	Measure	R	S	$R{+}S$	C	$R{+}C$	$R{+}S{+}C$
Qwen3-4B		0.511 / 0.527	0.536 / 0.533	0.533 / 0.545	0.564 / 0.586	0.566 / 0.590	0.564 / 0.578
n=1,029	Across / within [95% CI]	[0.499, 0.554]	[0.507, 0.557]	[0.519, 0.569]	[0.562, 0.611]	[0.565, 0.616]	[0.552, 0.604]
		0.040	0.329	0.343	0.786	0.815	0.747
	Conditional logit β [95% CI]	[−0.046, 0.126]	[0.145, 0.513]	[0.147, 0.538]	[0.613, 0.958]	[0.642, 0.988]	[0.580, 0.915]
Qwen3-8B		0.541 / 0.550	0.536 / 0.531	0.525 / 0.519	0.558 / 0.579	0.545 / 0.561	0.544 / 0.564
n=969	Across / within [95% CI]	[0.521, 0.578]	[0.505, 0.557]	[0.493, 0.545]	[0.552, 0.605]	[0.534, 0.587]	[0.537, 0.592]
		0.204	0.335	0.211	0.674	0.426	0.415
	Conditional logit β [95% CI]	[0.098, 0.309]	[0.182, 0.489]	[0.035, 0.387]	[0.514, 0.833]	[0.263, 0.589]	[0.253, 0.577]
Qwen3-14B		0.485 / 0.493	0.515 / 0.503	0.512 / 0.505	0.544 / 0.547	0.544 / 0.560	0.545 / 0.557
n=939	Across / within [95% CI]	[0.466, 0.521]	[0.477, 0.529]	[0.478, 0.530]	[0.520, 0.574]	[0.532, 0.586]	[0.530, 0.584]
		−0.032	0.079	0.064	0.566	0.567	0.563
	Conditional logit β [95% CI]	[−0.118, 0.054]	[−0.084, 0.242]	[−0.105, 0.234]	[0.391, 0.741]	[0.391, 0.743]	[0.392, 0.734]
gpt-oss-20b		0.512 / 0.530	0.536 / 0.526	0.533 / 0.515	0.558 / 0.566	0.559 / 0.559	0.559 / 0.560
n=1,389	Across / within [95% CI]	[0.505, 0.554]	[0.504, 0.548]	[0.494, 0.537]	[0.544, 0.587]	[0.537, 0.581]	[0.537, 0.581]
		0.740	0.167	0.132	0.392	0.409	0.376
	Conditional logit β [95% CI]	[0.472, 1.007]	[0.044, 0.290]	[0.008, 0.255]	[0.265, 0.519]	[0.281, 0.537]	[0.251, 0.500]
gpt-oss-120b		0.523 / 0.539	0.534 / 0.527	0.534 / 0.526	0.556 / 0.558	0.555 / 0.552	0.551 / 0.544
n=978	Across / within [95% CI]	[0.512, 0.567]	[0.501, 0.553]	[0.499, 0.553]	[0.533, 0.585]	[0.527, 0.579]	[0.519, 0.571]
		0.317	0.175	0.172	0.403	0.405	0.359
	Conditional logit β [95% CI]	[0.162, 0.472]	[0.034, 0.317]	[0.033, 0.311]	[0.255, 0.550]	[0.255, 0.554]	[0.210, 0.508]

Appendix Table A53: **Prompt-visible repeated-run evaluation under the primary surface-plus-Cohere monitor.** Entries give across-item AUROC / within-item AUROC, with 95% item-level bootstrap intervals for the within-item value. Conditional-logit entries give the coefficient for a one-SD increase in monitor score and its 95% interval. Both AUROCs use the same post-outcome discordant held-out MMLU-Pro items across three independent runs per item. Monitors are trained on the primary run. Across-item AUROC pools runs across items; within-item AUROC compares a correct and an incorrect run of the same item and model. The task prompt is supplied to every display and held fixed within each pair. There are 939–1,389 discordant items per model.

Prompt-visible displays					
Model	Measure	$q+R$	$q+R+S$	$q+R+C$	$q+R+S+C$
Qwen3-4B		0.510 / 0.517	0.524 / 0.543	0.562 / 0.580	0.563 / 0.584
n=1,029	Across / within [95% CI]	[0.488, 0.545]	[0.517, 0.568]	[0.555, 0.605]	[0.558, 0.608]
		0.864	0.574	0.814	0.799
	Conditional logit β [95% CI]	[0.093, 1.635]	[0.275, 0.874]	[0.639, 0.988]	[0.626, 0.972]
Qwen3-8B		0.523 / 0.524	0.531 / 0.547	0.548 / 0.561	0.548 / 0.567
n=969	Across / within [95% CI]	[0.496, 0.553]	[0.522, 0.573]	[0.535, 0.587]	[0.540, 0.594]
		0.494	0.365	0.455	0.473
	Conditional logit β [95% CI]	[0.184, 0.804]	[0.147, 0.583]	[0.289, 0.621]	[0.305, 0.640]
Qwen3-14B		0.509 / 0.507	0.510 / 0.486	0.547 / 0.555	0.543 / 0.552
n=939	Across / within [95% CI]	[0.476, 0.537]	[0.460, 0.513]	[0.529, 0.582]	[0.526, 0.579]
		0.175	−0.059	0.585	0.527
	Conditional logit β [95% CI]	[−1.049, 1.399]	[−0.307, 0.189]	[0.408, 0.763]	[0.354, 0.700]
gpt-oss-20b		0.514 / 0.485	0.529 / 0.536	0.558 / 0.554	0.558 / 0.560
n=1,389	Across / within [95% CI]	[0.460, 0.510]	[0.513, 0.558]	[0.532, 0.576]	[0.539, 0.582]
		−0.651	0.377	0.389	0.380
	Conditional logit β [95% CI]	[−2.004, 0.702]	[0.153, 0.600]	[0.258, 0.520]	[0.251, 0.509]
gpt-oss-120b		0.512 / 0.482	0.525 / 0.531	0.556 / 0.564	0.560 / 0.564
n=978	Across / within [95% CI]	[0.452, 0.512]	[0.505, 0.557]	[0.539, 0.591]	[0.538, 0.589]
		−0.736	0.279	0.459	0.472
	Conditional logit β [95% CI]	[−2.378, 0.906]	[0.020, 0.538]	[0.298, 0.620]	[0.315, 0.630]

Appendix Table A54: **Repeated-run signal under several controls for visible text length.** The upper panel reports AUROC when ranking across selected questions and when comparing runs of the same question, averaged equally over the five target models. The matched-length panels keep correct–incorrect run pairs whose trace word counts differ by no more than 20% or 10% of the longer trace. The aggregate summary–trace estimates are equal-model means; their intervals come from a paired item bootstrap stratified by model, with every eligible pair for each sampled item retained. Other brackets are 95% item-bootstrap intervals, and n is the exact number of eligible pairs. All across-item and within-item estimates use the same items with three runs and both observed outcomes.

Display / model	Primary	Words	Words + sentences	Length-adjusted	Within 20%	Within 10%
Across-item / within-item AUROC, equal-model mean						
Output-only displays						
R	0.515 / 0.528	0.494 / 0.491	0.506 / 0.503	0.514 / 0.524	–	–
S	0.531 / 0.524	0.524 / 0.514	0.522 / 0.515	0.514 / 0.508	–	–
$R{+}S$	0.528 / 0.522	0.524 / 0.518	0.516 / 0.507	0.521 / 0.519	–	–
C	0.556 / 0.567	0.539 / 0.553	0.539 / 0.545	0.527 / 0.538	–	–
$R{+}C$	0.554 / 0.564	0.539 / 0.552	0.539 / 0.546	0.524 / 0.528	–	–
$R{+}S{+}C$	0.553 / 0.561	0.539 / 0.553	0.540 / 0.547	0.524 / 0.529	–	–
Prompt-visible displays						
$q{+}R$	0.514 / 0.503	0.508 / 0.508	0.507 / 0.504	0.510 / 0.504	–	–
$q{+}R{+}S$	0.524 / 0.529	0.512 / 0.518	0.511 / 0.516	0.519 / 0.519	–	–
$q{+}R{+}C$	0.554 / 0.563	0.538 / 0.552	0.539 / 0.547	0.524 / 0.536	–	–
$q{+}R{+}S{+}C$	0.555 / 0.565	0.539 / 0.553	0.539 / 0.547	0.524 / 0.533	–	–

Matched-length summary–trace comparison, equal-model mean

Restriction	Summary	Trace	Trace minus summary [95% CI]	Eligible pairs
Within 20%	0.502	0.533	+0.032 [+0.009, +0.056]	3,412
Within 10%	0.495	0.532	+0.037 [+0.004, +0.069]	1,701

Trace-only display, primary reader: within-item AUROC [95% CI] (n eligible pairs)

Qwen3-4B	0.586 ($n = 2{,}058$)	–	–	–	0.539 [0.501, 0.578] ($n = 765$)	0.520 [0.467, 0.575] ($n = 371$)
Qwen3-8B	0.579 ($n = 1{,}938$)	–	–	–	0.517 [0.480, 0.556] ($n = 755$)	0.526 [0.476, 0.575] ($n = 406$)
Qwen3-14B	0.547 ($n = 1{,}878$)	–	–	–	0.481 [0.440, 0.519] ($n = 674$)	0.476 [0.419, 0.533] ($n = 317$)
gpt-oss-20b	0.566 ($n = 2{,}778$)	–	–	–	0.579 [0.537, 0.621] ($n = 605$)	0.599 [0.542, 0.652] ($n = 299$)
gpt-oss-120b	0.558 ($n = 1{,}956$)	–	–	–	0.551 [0.510, 0.593] ($n = 613$)	0.539 [0.482, 0.596] ($n = 308$)

Appendix Table A55: **A conclusion-region trace extract is compared with the model-written summary on repeated runs.** Entries give across-item AUROC / within-item AUROC; square brackets give the 95% item-bootstrap interval for within-item AUROC. Self-summary uses the original preservation-tuned summary, whereas Last90 deterministically places up to the final 90 whitespace-delimited trace words in the summary slot and therefore uses the entire trace when it is shorter. It is a conclusion-region-biased control rather than a pure ending-only test. Monitors use the same primary surface-plus-Cohere feature family, are trained on run 1, and score held-out MMLU-Pro items across three independent runs. Both AUROCs use the same post-outcome discordant items. Across-item AUROC pools runs across those items, while within-item AUROC compares correct and incorrect runs of one item.

Model	S self-summary	S Last90	$R{+}S$ self-summary	$R{+}S$ Last90
Qwen3-4B	0.536 / 0.533 [0.507, 0.557]	0.550 / 0.559 [0.534, 0.584]	0.533 / 0.545 [0.519, 0.569]	0.541 / 0.543 [0.518, 0.569]
Qwen3-8B	0.536 / 0.531 [0.505, 0.557]	0.558 / 0.544 [0.519, 0.570]	0.525 / 0.519 [0.493, 0.545]	0.528 / 0.542 [0.518, 0.566]
Qwen3-14B	0.515 / 0.503 [0.477, 0.529]	0.524 / 0.525 [0.498, 0.551]	0.512 / 0.505 [0.478, 0.530]	0.525 / 0.522 [0.496, 0.548]
gpt-oss-20b	0.536 / 0.526 [0.504, 0.548]	0.548 / 0.536 [0.514, 0.558]	0.533 / 0.515 [0.494, 0.537]	0.548 / 0.538 [0.516, 0.560]
gpt-oss-120b	0.534 / 0.527 [0.501, 0.553]	0.556 / 0.565 [0.539, 0.593]	0.534 / 0.526 [0.499, 0.553]	0.554 / 0.574 [0.547, 0.600]

Appendix Table A56: **Prompt-visible repeated-run evaluation holds the item fixed when measuring run-level signal.** Entries give across-item AUROC / within-item AUROC, with 95% item-level bootstrap intervals in square brackets for the within-item value. Conditional-logit entries give the coefficient for a one-SD increase in monitor score and its 95% interval. The fully open monitor uses surface features plus word- and character-level TF–IDF and is trained on the primary run. Both AUROCs use the same post-outcome discordant items. Across-item AUROC pools held-out runs across those items. Within-item AUROC compares a correct and an incorrect run of the same MMLU-Pro item under the same model across three independent runs. The task prompt is supplied to every display and held fixed by the within-item pairing. For the two gpt-oss $q+R$ estimates, scores vary almost entirely between items, leaving little variation after the item fixed effects. Their large conditional-logit coefficients are therefore poorly scaled and are not used to support the main conclusions.

Prompt-visible displays					
Model	Measure	$q+R$	$q+R+S$	$q+R+C$	$q+R+S+C$
Qwen3-4B		0.519 / 0.535	0.550 / 0.563	0.590 / 0.606	0.589 / 0.607
n=1,029	Across / within [95% CI]	[0.506, 0.564]	[0.537, 0.588]	[0.581, 0.631]	[0.582, 0.632]
	Conditional logit β [95% CI]	0.441 [0.076, 0.806]	0.626 [0.447, 0.806]	0.756 [0.611, 0.902]	0.757 [0.611, 0.902]
Qwen3-8B		0.529 / 0.554	0.547 / 0.555	0.564 / 0.588	0.563 / 0.585
n=969	Across / within [95% CI]	[0.529, 0.580]	[0.528, 0.581]	[0.561, 0.614]	[0.558, 0.611]
	Conditional logit β [95% CI]	0.316 [0.131, 0.501]	0.375 [0.229, 0.522]	0.559 [0.419, 0.698]	0.554 [0.416, 0.693]
Qwen3-14B		0.508 / 0.520	0.529 / 0.544	0.558 / 0.571	0.558 / 0.571
n=939	Across / within [95% CI]	[0.488, 0.551]	[0.517, 0.570]	[0.545, 0.597]	[0.544, 0.597]
	Conditional logit β [95% CI]	0.555 [-0.522, 1.632]	0.267 [0.092, 0.442]	0.512 [0.362, 0.661]	0.506 [0.357, 0.654]
gpt-oss-20b		0.520 / 0.518	0.551 / 0.543	0.566 / 0.557	0.568 / 0.562
n=1,389	Across / within [95% CI]	[0.494, 0.543]	[0.521, 0.566]	[0.535, 0.579]	[0.539, 0.583]
	Conditional logit β [95% CI]	4.651 [0.441, 8.860]	0.362 [0.221, 0.504]	0.372 [0.251, 0.493]	0.390 [0.268, 0.512]
gpt-oss-120b		0.505 / 0.539	0.543 / 0.572	0.561 / 0.568	0.561 / 0.570
n=978	Across / within [95% CI]	[0.509, 0.570]	[0.546, 0.597]	[0.542, 0.595]	[0.544, 0.596]
	Conditional logit β [95% CI]	6.238 [2.521, 9.955]	0.440 [0.277, 0.603]	0.442 [0.298, 0.587]	0.452 [0.309, 0.595]

Appendix Table A57: **A conclusion-region trace extract is compared with the model-written summary on repeated runs.** Entries give across-item AUROC / within-item AUROC; square brackets give the 95% item-bootstrap interval for within-item AUROC. Self-summary uses the original preservation-tuned summary, whereas Last90 deterministically places up to the final 90 whitespace-delimited trace words in the summary slot and therefore uses the entire trace when it is shorter. It is a conclusion-region-biased control rather than a pure ending-only test. Monitors use the fully open surface-plus-TF–IDF feature family, are trained on run 1, and score held-out MMLU-Pro items across three independent runs. Both AUROCs use the same post-outcome discordant items. Across-item AUROC pools runs across those items, while within-item AUROC compares correct and incorrect runs of one item.

Model	S self-summary	S Last90	$R{+}S$ self-summary	$R{+}S$ Last90
Qwen3-4B	0.575 / 0.565 [0.539, 0.590]	0.597 / 0.592 [0.566, 0.617]	0.574 / 0.568 [0.542, 0.592]	0.596 / 0.589 [0.563, 0.614]
Qwen3-8B	0.563 / 0.569 [0.543, 0.595]	0.580 / 0.554 [0.528, 0.581]	0.547 / 0.551 [0.524, 0.578]	0.553 / 0.544 [0.518, 0.571]
Qwen3-14B	0.543 / 0.543 [0.515, 0.569]	0.563 / 0.562 [0.536, 0.588]	0.546 / 0.542 [0.513, 0.568]	0.561 / 0.559 [0.533, 0.585]
gpt-oss-20b	0.561 / 0.565 [0.543, 0.587]	0.575 / 0.558 [0.536, 0.580]	0.562 / 0.571 [0.549, 0.593]	0.576 / 0.558 [0.536, 0.581]
gpt-oss-120b	0.565 / 0.567 [0.542, 0.593]	0.588 / 0.596 [0.571, 0.622]	0.566 / 0.572 [0.546, 0.597]	0.589 / 0.596 [0.570, 0.622]

Appendix Table A58: **Summary-based pairwise monitors remain near chance; trace-based monitors are modest.** Pairwise logistic monitors are trained on correct-minus-incorrect feature differences within discordant MMLU-Pro training items, select regularization on validation items, and are evaluated on held-out discordant items. Values are mean ± SD across data splits, with descriptive 2.5th–97.5th percentiles of the five estimates in square brackets. These percentiles summarize split sensitivity and are not confidence intervals.

Model	R	S	$R{+}S$	C	$R{+}C$	$R{+}S{+}C$
Qwen3-4B	0.511 ± 0.020 [0.494, 0.542]	0.533 ± 0.013 [0.523, 0.552]	0.535 ± 0.023 [0.510, 0.558]	0.562 ± 0.028 [0.524, 0.586]	0.580 ± 0.018 [0.562, 0.601]	0.556 ± 0.017 [0.543, 0.580]
Qwen3-8B	0.530 ± 0.023 [0.500, 0.554]	0.514 ± 0.037 [0.474, 0.546]	0.524 ± 0.019 [0.502, 0.548]	0.558 ± 0.029 [0.521, 0.593]	0.544 ± 0.029 [0.511, 0.570]	0.564 ± 0.016 [0.546, 0.581]
Qwen3-14B	0.532 ± 0.035 [0.495, 0.579]	0.512 ± 0.019 [0.483, 0.525]	0.518 ± 0.010 [0.507, 0.531]	0.543 ± 0.015 [0.520, 0.554]	0.553 ± 0.017 [0.529, 0.568]	0.543 ± 0.011 [0.534, 0.556]
gpt-oss-20b	0.536 ± 0.014 [0.525, 0.551]	0.522 ± 0.013 [0.508, 0.536]	0.536 ± 0.011 [0.525, 0.552]	0.544 ± 0.015 [0.527, 0.565]	0.534 ± 0.022 [0.515, 0.565]	0.551 ± 0.022 [0.525, 0.569]
gpt-oss-120b	0.541 ± 0.031 [0.508, 0.581]	0.528 ± 0.014 [0.516, 0.549]	0.542 ± 0.042 [0.506, 0.605]	0.572 ± 0.016 [0.553, 0.592]	0.577 ± 0.022 [0.549, 0.601]	0.561 ± 0.031 [0.512, 0.582]

Appendix Table A59: **Summary-based pairwise monitors remain near chance; trace-based monitors are modest.** The primary monitor uses rate-based surface features plus Cohere embeddings. The task prompt is included in every display. Pairwise logistic monitors are trained on correct-minus-incorrect feature differences within discordant MMLU-Pro training items, select regularization on validation items, and are evaluated on held-out discordant items. Values are mean ± SD across data splits, with descriptive 2.5th–97.5th percentiles of the five estimates in square brackets. These percentiles summarize split sensitivity and are not confidence intervals.

Model	$q{+}R$	$q{+}R{+}S$	$q{+}R{+}C$	$q{+}R{+}S{+}C$
Qwen3-4B	0.530 ± 0.020 [0.504, 0.553]	0.555 ± 0.015 [0.540, 0.574]	0.555 ± 0.012 [0.541, 0.568]	0.566 ± 0.016 [0.545, 0.586]
Qwen3-8B	0.515 ± 0.013 [0.499, 0.531]	0.544 ± 0.019 [0.520, 0.565]	0.559 ± 0.012 [0.544, 0.573]	0.565 ± 0.019 [0.546, 0.591]
Qwen3-14B	0.510 ± 0.033 [0.478, 0.551]	0.514 ± 0.035 [0.481, 0.559]	0.570 ± 0.011 [0.558, 0.585]	0.550 ± 0.027 [0.508, 0.572]
gpt-oss-20b	0.501 ± 0.022 [0.477, 0.530]	0.543 ± 0.017 [0.529, 0.565]	0.538 ± 0.025 [0.510, 0.562]	0.547 ± 0.019 [0.532, 0.569]
gpt-oss-120b	0.522 ± 0.026 [0.499, 0.554]	0.548 ± 0.024 [0.522, 0.583]	0.576 ± 0.021 [0.545, 0.595]	0.586 ± 0.014 [0.566, 0.599]

Appendix Table A60: **Direct item-conditional repeated-run training compares correct and incorrect runs of the same item.** Pairwise logistic monitors are trained on correct-minus-incorrect feature differences within discordant MMLU-Pro training items, select regularization on validation items, and are evaluated on held-out discordant items. The fully open specification uses surface features plus word- and character-level TF–IDF. Values are mean ± SD across data splits, with descriptive 2.5th–97.5th percentiles of the five estimates in square brackets. These percentiles summarize split sensitivity and are not confidence intervals.

Model	$q+R$	$q+R+S$	$q+R+C$	$q+R+S+C$
Qwen3-4B	0.510 ± 0.021 [0.487, 0.541]	0.562 ± 0.022 [0.537, 0.586]	0.588 ± 0.017 [0.572, 0.613]	0.592 ± 0.010 [0.580, 0.605]
Qwen3-8B	0.536 ± 0.022 [0.512, 0.558]	0.534 ± 0.029 [0.494, 0.568]	0.570 ± 0.006 [0.561, 0.575]	0.572 ± 0.009 [0.562, 0.583]
Qwen3-14B	0.525 ± 0.019 [0.503, 0.550]	0.507 ± 0.037 [0.468, 0.557]	0.567 ± 0.037 [0.521, 0.611]	0.571 ± 0.044 [0.517, 0.623]
gpt-oss-20b	0.516 ± 0.024 [0.491, 0.552]	0.544 ± 0.015 [0.524, 0.558]	0.555 ± 0.021 [0.532, 0.580]	0.562 ± 0.019 [0.534, 0.577]
gpt-oss-120b	0.549 ± 0.029 [0.514, 0.586]	0.531 ± 0.025 [0.495, 0.553]	0.564 ± 0.020 [0.553, 0.596]	0.566 ± 0.009 [0.553, 0.573]

Appendix Table A61: **The summary–trace ordering persists in both discordance-composition strata, but trace discrimination is concentrated among items with exactly one incorrect run.** The upper block distinguishes all outcome-discordant MMLU-Pro items among the 12,032 items with three graded runs per model from the analyzed subset that appeared in the test partition for at least one of five data splits. Parentheses give the analyzed share of all outcome-discordant items. The lower block pools the five target models and reports within-item AUROC for items with exactly one incorrect run (n=2,829), exactly two incorrect runs (n=2,475), and both strata combined. Square brackets give 95% intervals from 5,000 item-bootstrap draws that resample model–item units within target model. The model-balanced pooled estimate assigns every analyzed item for model m weight $1/n_m$, so the five analyzed model samples contribute equally. This reweighting does not correct selection into outcome discordance or held-out test inclusion, and it does not score concordant or unanalyzed items.

Model	**All three-run items**		**Analyzed discordant items**		
	Three graded	All discordant	1 incorrect	2 incorrect	Total (share)
Qwen3-4B	12,032	1,530	526	503	1,029 (67.3%)
Qwen3-8B	12,032	1,481	507	462	969 (65.4%)
Qwen3-14B	12,032	1,364	498	441	939 (68.8%)
gpt-oss-20b	12,032	2,062	761	628	1,389 (67.4%)
gpt-oss-120b	12,032	1,406	537	441	978 (69.6%)

Within-item AUROC by discordance composition					
Display	1 incorrect	2 incorrect	Pooled	Model-balanced pooled	Within pairs
Prompt-withheld displays					
R	0.525 [0.508, 0.542]	0.532 [0.515, 0.550]	0.528 [0.516, 0.541]	0.528 [0.516, 0.540]	10,608
S	0.551 [0.535, 0.567]	0.494 [0.478, 0.511]	0.524 [0.513, 0.536]	0.524 [0.513, 0.536]	10,608
R+S	0.548 [0.532, 0.563]	0.493 [0.476, 0.509]	0.522 [0.511, 0.533]	0.522 [0.511, 0.533]	10,608
C	0.611 [0.596, 0.626]	0.518 [0.502, 0.535]	0.568 [0.556, 0.579]	0.567 [0.556, 0.579]	10,608
R+C	0.605 [0.590, 0.621]	0.518 [0.502, 0.534]	0.564 [0.553, 0.575]	0.564 [0.553, 0.576]	10,608
R+S+C	0.604 [0.588, 0.619]	0.512 [0.495, 0.527]	0.561 [0.549, 0.572]	0.561 [0.549, 0.572]	10,608
Prompt-visible displays					
q+R	0.504 [0.487, 0.522]	0.499 [0.481, 0.517]	0.502 [0.489, 0.514]	0.503 [0.490, 0.516]	10,608
q+R+S	0.553 [0.537, 0.568]	0.503 [0.486, 0.519]	0.530 [0.518, 0.541]	0.529 [0.517, 0.540]	10,608
q+R+C	0.605 [0.589, 0.619]	0.514 [0.497, 0.530]	0.562 [0.551, 0.573]	0.563 [0.551, 0.574]	10,608
q+R+S+C	0.608 [0.593, 0.623]	0.517 [0.500, 0.533]	0.565 [0.554, 0.577]	0.565 [0.554, 0.577]	10,608

Appendix Table A62: **The repeated-run subset is compared with held-out items whose outcomes remain stable.** Discordant items have both correct and incorrect outcomes across three runs; concordant items have one outcome across all runs. Prompt score is the held-out probability from the primary prompt-only reader, averaged over available data splits. Trace length is the median [25th, 75th percentile] number of words. Replay confidence is the target model's maximum answer-choice probability under provider-side replay. The domain panel reports total variation (TV) between the two domain distributions and the largest positive percentage-point difference for discordant items.

Model	Group	n	Run 1 accuracy	Prompt score	Trace words	Replay confidence
Qwen3-4B	Concordant	7,097	0.741	0.627	863 [468, 1,782]	0.9948
	Discordant	1,029	0.507	0.561	2,158 [1,157, 4,135]	0.9831
Qwen3-8B	Concordant	7,088	0.787	0.606	979 [527, 2,012]	0.9531
	Discordant	969	0.518	0.544	2,527 [1,419, 4,414]	0.9484
Qwen3-14B	Concordant	7,163	0.812	0.696	719 [415, 1,522.5]	0.9943
	Discordant	939	0.524	0.613	1,905 [1,077, 3,266.5]	0.9871
gpt-oss-20b	Concordant	6,734	0.794	0.630	200 [94, 445]	0.9997
	Discordant	1,389	0.508	0.554	591 [319, 1,092]	0.9997
gpt-oss-120b	Concordant	7,130	0.829	0.650	138 [74, 268]	0.9999
	Discordant	978	0.530	0.576	379 [224.2, 638.8]	1.0000

Domain composition

Model	TV distance	Largest discordant excess	Difference
Qwen3-4B	17.3%	law	+5.9 pp
Qwen3-8B	21.4%	law	+8.9 pp
Qwen3-14B	26.4%	law	+12.0 pp
gpt-oss-20b	22.8%	law	+11.6 pp
gpt-oss-120b	22.9%	law	+13.5 pp

Appendix Table A63: **The fixed-prompt GPT-5-mini reader is evaluated on all ten tested combinations of the response, summary, trace, and prompt.** Across-item AUROC ranks runs across questions; within-item AUROC compares runs of the same question. Both estimates use the same 1,389 analyzed outcome-discordant MMLU-Pro questions, three gpt-oss-20b runs per item, and the fixed-prompt `gpt-5-mini-2025-08-07` reader. Brackets give paired 95% item-bootstrap intervals (5,000 resamples). Each comparison subtracts the second named display from the first on every paired bootstrap draw.

Display or comparison	Across-item estimate [95% CI]	Within-item estimate [95% CI]
Display AUROC		
Response (R)	0.501 [0.485, 0.517]	0.503 [0.486, 0.521]
Summary (S)	0.673 [0.655, 0.690]	0.649 [0.628, 0.669]
Reasoning trace (C)	0.682 [0.667, 0.698]	0.683 [0.663, 0.702]
Response + summary	0.676 [0.659, 0.694]	0.656 [0.635, 0.677]
Response + reasoning trace	0.700 [0.684, 0.716]	0.690 [0.670, 0.710]
Response + summary + reasoning trace	0.704 [0.688, 0.721]	0.693 [0.673, 0.713]
Prompt + response	0.834 [0.817, 0.850]	0.800 [0.781, 0.818]
Prompt + response + summary	0.806 [0.789, 0.823]	0.776 [0.757, 0.795]
Prompt + response + reasoning trace	0.748 [0.732, 0.764]	0.736 [0.717, 0.756]
Prompt + response + summary + reasoning trace	0.746 [0.729, 0.762]	0.721 [0.701, 0.741]
Paired AUROC differences		
Summary minus response	+0.172 [+0.149, +0.195]	+0.146 [+0.118, +0.172]
Reasoning trace minus summary	+0.009 [-0.008, +0.027]	+0.034 [+0.008, +0.059]
Summary added to response	+0.175 [+0.151, +0.199]	+0.153 [+0.125, +0.180]
Reasoning trace versus summary after response	+0.024 [+0.006, +0.041]	+0.034 [+0.008, +0.057]
Summary added to prompt + response	-0.028 [-0.039, -0.017]	-0.024 [-0.040, -0.007]
Reasoning trace added to prompt + response	-0.086 [-0.100, -0.072]	-0.064 [-0.084, -0.044]
Summary added after prompt + response + reasoning trace	-0.002 [-0.012, +0.008]	-0.015 [-0.032, +0.002]

Appendix Table A64: **Provider-only replay features add a mean hidden-state gain that survives prompt conditioning.** Held-out test AUROC by setting, averaged over five data splits (± SD across data splits). $+Z_{\text{prob}}$, $+Z_{\text{hidden}}$, and $+Z$ add replay-derived answer-probability, raw hidden-state, and combined replay features to the full public display. The three rightmost columns repeat the comparison on top of the prompt-conditioned public display $q{+}R{+}S{+}C$. The final row averages per-setting means across all 15 settings (± between-setting SD).

Benchmark	Model	$R{+}S{+}C$	$+Z_{\text{prob}}$	$+Z_{\text{hidden}}$	$+Z$	$q{+}R{+}S{+}C$	$q\ldots+Z_{\text{hidden}}$	$q\ldots+Z$
GPQA-Diamond	Qwen3-4B	0.732 ± 0.067	0.705 ± 0.079	0.738 ± 0.070	0.744 ± 0.059	0.725 ± 0.084	0.739 ± 0.065	0.737 ± 0.067
GPQA-Diamond	Qwen3-8B	0.839 ± 0.062	0.839 ± 0.063	0.858 ± 0.034	0.859 ± 0.038	0.836 ± 0.058	0.853 ± 0.036	0.855 ± 0.033
GPQA-Diamond	Qwen3-14B	0.800 ± 0.058	0.773 ± 0.072	0.774 ± 0.080	0.767 ± 0.078	0.758 ± 0.056	0.765 ± 0.084	0.759 ± 0.084
GPQA-Diamond	gpt-oss-20b	0.707 ± 0.096	0.698 ± 0.098	0.771 ± 0.075	0.749 ± 0.103	0.723 ± 0.077	0.769 ± 0.083	0.751 ± 0.109
GPQA-Diamond	gpt-oss-120b	0.770 ± 0.057	0.771 ± 0.066	0.798 ± 0.069	0.799 ± 0.069	0.747 ± 0.059	0.795 ± 0.071	0.796 ± 0.071
MMLU-Pro	Qwen3-4B	0.823 ± 0.010	0.837 ± 0.009	0.859 ± 0.008	0.859 ± 0.008	0.824 ± 0.009	0.859 ± 0.007	0.859 ± 0.007
MMLU-Pro	Qwen3-8B	0.821 ± 0.004	0.831 ± 0.007	0.861 ± 0.007	0.861 ± 0.007	0.822 ± 0.004	0.861 ± 0.007	0.861 ± 0.007
MMLU-Pro	Qwen3-14B	0.825 ± 0.007	0.844 ± 0.007	0.866 ± 0.004	0.865 ± 0.003	0.826 ± 0.006	0.866 ± 0.004	0.866 ± 0.004
MMLU-Pro	gpt-oss-20b	0.803 ± 0.015	0.808 ± 0.015	0.844 ± 0.011	0.844 ± 0.011	0.804 ± 0.015	0.845 ± 0.011	0.845 ± 0.011
MMLU-Pro	gpt-oss-120b	0.810 ± 0.005	0.814 ± 0.005	0.845 ± 0.005	0.845 ± 0.005	0.809 ± 0.004	0.845 ± 0.005	0.845 ± 0.005
Omni-MATH-2-Filtered	Qwen3-4B	0.870 ± 0.015	0.870 ± 0.016	0.893 ± 0.015	0.894 ± 0.013	0.870 ± 0.015	0.893 ± 0.015	0.894 ± 0.013
Omni-MATH-2-Filtered	Qwen3-8B	0.875 ± 0.022	0.878 ± 0.021	0.912 ± 0.012	0.912 ± 0.012	0.870 ± 0.017	0.912 ± 0.011	0.912 ± 0.011
Omni-MATH-2-Filtered	Qwen3-14B	0.876 ± 0.008	0.877 ± 0.008	0.891 ± 0.005	0.891 ± 0.005	0.871 ± 0.006	0.890 ± 0.005	0.890 ± 0.005
Omni-MATH-2-Filtered	gpt-oss-20b	0.824 ± 0.022	0.825 ± 0.021	0.864 ± 0.020	0.864 ± 0.020	0.824 ± 0.016	0.865 ± 0.020	0.865 ± 0.021
Omni-MATH-2-Filtered	gpt-oss-120b	0.832 ± 0.019	0.835 ± 0.019	0.849 ± 0.011	0.849 ± 0.011	0.832 ± 0.016	0.848 ± 0.010	0.848 ± 0.010
Mean		0.814 ± 0.048	0.814 ± 0.056	0.841 ± 0.050	0.840 ± 0.052	0.809 ± 0.050	0.840 ± 0.050	0.839 ± 0.053

Appendix Table A65: **The mean hidden-state gain is largest with the run's own trace and is lost under shuffled replay.** Held-out test AUROC of the $R{+}S{+}C{+}Z_{\text{hidden}}$ monitor when the hidden state is extracted from replays of different contexts: the standard task prompt plus the run's reasoning trace ($q{+}C$), the prompt alone, the trace alone, the prompt plus final response, the prompt plus the aggressively masked trace, and the prompt plus a trace from a different item, reassigned using a fixed random seed. The public $R{+}S{+}C$ column repeats the comparison without replay. This pattern holds on average over the reported subset and need not hold in every setting. Values average five data splits (± SD across data splits) for all five models on GPQA-Diamond and Qwen3-8B on MMLU-Pro and Omni-MATH-2-Filtered.

Benchmark	Model	$R{+}S{+}C$	$q{+}C$	q only	C only	$q{+}R$	q+masked C	q+shuffled C
GPQA-Diamond	gpt-oss-120b	0.770 ± 0.057	0.798 ± 0.069	0.767 ± 0.077	0.830 ± 0.058	0.789 ± 0.064	0.713 ± 0.044	0.672 ± 0.144
GPQA-Diamond	gpt-oss-20b	0.707 ± 0.096	0.771 ± 0.075	0.693 ± 0.094	0.730 ± 0.089	0.658 ± 0.079	0.708 ± 0.062	0.674 ± 0.063
GPQA-Diamond	Qwen3-14B	0.800 ± 0.058	0.774 ± 0.080	0.724 ± 0.102	0.797 ± 0.086	0.758 ± 0.132	0.780 ± 0.102	0.721 ± 0.059
GPQA-Diamond	Qwen3-4B	0.732 ± 0.067	0.738 ± 0.070	0.660 ± 0.080	0.698 ± 0.103	0.656 ± 0.041	0.746 ± 0.069	0.685 ± 0.032
GPQA-Diamond	Qwen3-8B	0.839 ± 0.062	0.858 ± 0.034	0.828 ± 0.060	0.849 ± 0.038	0.754 ± 0.056	0.854 ± 0.035	0.829 ± 0.052
MMLU-Pro	Qwen3-8B	0.821 ± 0.004	0.861 ± 0.007	0.833 ± 0.010	0.857 ± 0.008	0.842 ± 0.005	0.857 ± 0.007	0.822 ± 0.006
Omni-MATH-2-Filtered	Qwen3-8B	0.875 ± 0.022	0.912 ± 0.012	0.874 ± 0.019	0.907 ± 0.010	0.908 ± 0.014	0.914 ± 0.010	0.866 ± 0.018

Appendix Table A66: **Availability of final-answer extraction and replay features.** "Answer extracted" is the percentage of items whose final answer was extracted successfully; "Extraction failures" gives the corresponding count. "Replay complete" is the share with a completed prompt-conditioned calculation used to obtain answer probabilities and hidden states. "Within context window" is the share of inputs that fit the model's context window. "Usable replay" requires both a completed replay and an input within the context window. "Hidden-state vectors" gives the number of items with an available final-token vector.

Benchmark	Model	Items	Answer extracted	Extraction failures	Replay complete	Within context window	Usable replay	Hidden-state vectors
GPQA-Diamond	Qwen3-4B	198	100%	0	100%	100%	100%	198
GPQA-Diamond	Qwen3-8B	198	100%	0	100%	100%	100%	198
GPQA-Diamond	Qwen3-14B	198	100%	0	100%	100%	100%	198
GPQA-Diamond	gpt-oss-20b	198	100%	0	100%	100%	100%	198
GPQA-Diamond	gpt-oss-120b	198	100%	0	100%	100%	100%	198
MMLU-Pro	Qwen3-4B	12,032	100%	1	100%	100%	100%	12,032
MMLU-Pro	Qwen3-8B	12,032	100%	4	100%	100%	100%	12,032
MMLU-Pro	Qwen3-14B	12,032	100%	0	100%	100%	100%	12,032
MMLU-Pro	gpt-oss-20b	12,032	100%	5	100%	100%	100%	12,032
MMLU-Pro	gpt-oss-120b	12,032	100%	3	100%	100%	100%	12,032
Omni-MATH-2-Filtered	Qwen3-4B	4,181	95%	189	100%	100%	100%	4,181
Omni-MATH-2-Filtered	Qwen3-8B	4,181	95%	194	100%	100%	100%	4,181
Omni-MATH-2-Filtered	Qwen3-14B	4,181	95%	207	100%	100%	100%	4,181
Omni-MATH-2-Filtered	gpt-oss-20b	4,181	96%	152	100%	100%	100%	4,181
Omni-MATH-2-Filtered	gpt-oss-120b	4,181	97%	118	100%	100%	100%	4,181

Appendix Table A67: **Public monitors transfer across target models better than across benchmarks.** Held-out test AUROC on the target setting for probes trained on the source setting(s), averaged over five target data splits (± SD across data splits). Transfers reuse the public monitor features. No probe parameters are refit on the target setting. Within-benchmark comparisons transfer across models on Omni-MATH-2-Filtered; source runs follow the target split so target-test item identifiers are excluded from source train and validation. Cross-benchmark comparisons transfer the same model between MMLU-Pro and Omni-MATH-2-Filtered.

Source	Target	R	$R+S$	$R+C$	$R+S+C$	$q+R+S+C$
Within-benchmark, cross-model						
Omni-MATH-2-Filtered / gpt-oss-20b + gpt-oss-120b	Omni-MATH-2-Filtered / Qwen3-14B	0.702 ± 0.020	0.780 ± 0.006	0.778 ± 0.009	0.807 ± 0.005	0.801 ± 0.006
Omni-MATH-2-Filtered / Qwen3-8B	Omni-MATH-2-Filtered / Qwen3-14B	0.759 ± 0.011	0.814 ± 0.004	0.841 ± 0.011	0.842 ± 0.010	0.840 ± 0.006
Omni-MATH-2-Filtered / Qwen3-4B + Qwen3-8B + Qwen3-14B	Omni-MATH-2-Filtered / gpt-oss-120b	0.659 ± 0.017	0.741 ± 0.005	0.778 ± 0.016	0.787 ± 0.015	0.790 ± 0.014
Omni-MATH-2-Filtered / Qwen3-4B + Qwen3-8B + Qwen3-14B	Omni-MATH-2-Filtered / gpt-oss-20b	0.680 ± 0.022	0.732 ± 0.016	0.781 ± 0.018	0.783 ± 0.016	0.792 ± 0.011
Omni-MATH-2-Filtered / Qwen3-4B + Qwen3-8B	Omni-MATH-2-Filtered / Qwen3-14B	0.765 ± 0.016	0.817 ± 0.009	0.860 ± 0.005	0.866 ± 0.005	0.860 ± 0.002
Omni-MATH-2-Filtered / Qwen3-4B	Omni-MATH-2-Filtered / Qwen3-14B	0.748 ± 0.017	0.806 ± 0.011	0.845 ± 0.003	0.873 ± 0.002	0.849 ± 0.006
Mean (6 transfers)		0.719 ± 0.045	0.782 ± 0.038	0.814 ± 0.039	0.826 ± 0.040	0.822 ± 0.031
Cross-benchmark, same-model						
MMLU-Pro / gpt-oss-120b	Omni-MATH-2-Filtered / gpt-oss-120b	0.401 ± 0.010	0.719 ± 0.026	0.805 ± 0.023	0.813 ± 0.025	0.800 ± 0.026
MMLU-Pro / gpt-oss-20b	Omni-MATH-2-Filtered / gpt-oss-20b	0.618 ± 0.027	0.705 ± 0.020	0.789 ± 0.019	0.793 ± 0.019	0.796 ± 0.021
MMLU-Pro / Qwen3-14B	Omni-MATH-2-Filtered / Qwen3-14B	0.557 ± 0.043	0.685 ± 0.010	0.768 ± 0.012	0.819 ± 0.008	0.805 ± 0.007
MMLU-Pro / Qwen3-4B	Omni-MATH-2-Filtered / Qwen3-4B	0.646 ± 0.025	0.706 ± 0.016	0.853 ± 0.010	0.852 ± 0.013	0.854 ± 0.012
MMLU-Pro / Qwen3-8B	Omni-MATH-2-Filtered / Qwen3-8B	0.735 ± 0.026	0.713 ± 0.026	0.603 ± 0.014	0.642 ± 0.016	0.624 ± 0.010
Omni-MATH-2-Filtered / gpt-oss-120b	MMLU-Pro / gpt-oss-120b	0.526 ± 0.015	0.666 ± 0.013	0.735 ± 0.008	0.750 ± 0.010	0.745 ± 0.004
Omni-MATH-2-Filtered / gpt-oss-20b	MMLU-Pro / gpt-oss-20b	0.502 ± 0.012	0.643 ± 0.015	0.735 ± 0.016	0.743 ± 0.015	0.738 ± 0.015
Omni-MATH-2-Filtered / Qwen3-14B	MMLU-Pro / Qwen3-14B	0.490 ± 0.012	0.650 ± 0.013	0.696 ± 0.012	0.716 ± 0.013	0.727 ± 0.007
Omni-MATH-2-Filtered / Qwen3-4B	MMLU-Pro / Qwen3-4B	0.504 ± 0.019	0.638 ± 0.007	0.725 ± 0.008	0.721 ± 0.007	0.730 ± 0.007
Omni-MATH-2-Filtered / Qwen3-8B	MMLU-Pro / Qwen3-8B	0.573 ± 0.018	0.670 ± 0.011	0.645 ± 0.009	0.671 ± 0.012	0.659 ± 0.009
Mean (10 transfers)		0.555 ± 0.094	0.680 ± 0.030	0.735 ± 0.074	0.752 ± 0.067	0.748 ± 0.070

Appendix Table A68: **Higher-access monitors concentrate errors among lower-ranked runs.** Error rate among retained runs when the monitor keeps the indicated fraction of test runs ranked by predicted correctness and excludes the rest, averaged over data splits and the 15 settings. Coverage 1.0 is the unconditional error rate.

Level	50%	70%	80%	90%	100%
R	0.243	0.266	0.277	0.297	0.316
$R{+}S$	0.139	0.203	0.240	0.277	0.316
$R{+}S{+}C$	0.114	0.185	0.227	0.274	0.316
$R{+}S{+}C{+}Z$	0.098	0.168	0.217	0.267	0.316

Appendix Table A69: **Counts of extraction problems among completed runs.** Counts are taken from the saved primary runs before the data are divided for classifier training and evaluation. Empty trace means that the extracted reasoning trace contains no non-whitespace characters. No closing reasoning marker corresponds to `missing_think_close`; no answer extracted corresponds to `no_answer_found`. Across the 82,055 primary summaries, mean length is 77.0 words, 74.4% contain at most 90 words, and the maximum is 231 words.

Benchmark	Model	Runs	Empty trace	No closing reasoning marker	No answer extracted
GPQA-Diamond	Qwen3-4B	198	0	0	0
GPQA-Diamond	Qwen3-8B	198	3	0	0
GPQA-Diamond	Qwen3-14B	198	0	0	0
GPQA-Diamond	gpt-oss-20b	198	0	0	0
GPQA-Diamond	gpt-oss-120b	198	0	0	0
MMLU-Pro	Qwen3-4B	12,032	0	1	0
MMLU-Pro	Qwen3-8B	12,032	79	3	1
MMLU-Pro	Qwen3-14B	12,032	5	0	0
MMLU-Pro	gpt-oss-20b	12,032	0	0	5
MMLU-Pro	gpt-oss-120b	12,032	0	0	3
Omni-MATH-2-Filtered	Qwen3-4B	4,181	0	166	23
Omni-MATH-2-Filtered	Qwen3-8B	4,181	980	170	24
Omni-MATH-2-Filtered	Qwen3-14B	4,181	223	142	65
Omni-MATH-2-Filtered	gpt-oss-20b	4,181	0	0	151
Omni-MATH-2-Filtered	gpt-oss-120b	4,181	0	0	118

Appendix Table A70: **Full-public monitor discrimination by run stratum.** Held-out test AUROC of the $R+S+C$ public monitor within run strata, pooled over data splits and models per benchmark (test-set run counts in parentheses; sample-size weighted). Strata with fewer than 20 runs per setting or only one correctness class are excluded because their within-stratum AUROC is undefined. The missing-think-close category therefore contains Qwen3-8B only; the corresponding Qwen3-4B and Qwen3-14B strata are single-class. Length bins are within-setting terciles.

Stratum	Value	GPQA-Diamond	MMLU-Pro	Omni-MATH-2-Filtered
Trace length	long	0.632 (349)	0.750 (20100)	0.727 (6927)
Trace length	medium	0.698 (326)	0.795 (19992)	0.703 (6984)
Trace length	short	0.635 (325)	0.765 (20083)	0.853 (7014)
Summary length	long	0.694 (329)	0.771 (20035)	0.774 (6942)
Summary length	medium	0.736 (338)	0.798 (20154)	0.806 (6931)
Summary length	short	0.754 (333)	0.810 (19986)	0.880 (7052)
Trace self-correction	absent	—	0.721 (6914)	0.749 (1950)
Trace self-correction	present	0.741 (978)	0.803 (53261)	0.843 (18962)
Parser status	missing think close	—	—	0.727 (176)
Parser status	no answer found	—	—	0.737 (405)
Parser status	ok	0.745 (1000)	0.813 (60166)	0.848 (20034)
Choice count	10	—	0.821 (49831)	—
Choice count	3	—	0.821 (71)	—
Choice count	4	0.745 (1000)	0.739 (3076)	—
Choice count	5	—	0.671 (242)	—
Choice count	6	—	0.761 (485)	—
Choice count	7	—	0.771 (822)	—
Choice count	8	—	0.798 (1605)	—
Choice count	9	—	0.782 (4014)	—

Appendix Table A71: **Response-only discrimination varies by benchmark, model, and response format.** Columns report the short-response share (≤ 30 characters), length-only AUROC, selected-answer baseline, and the R monitor on all, short-only, and non-empty responses (mean over five data splits $\pm$ SD). Selected-answer and response-length baselines do not by themselves identify what drives the fitted monitor. A dash indicates that the length-only baseline is undefined.

Benchmark	Model	Short share	Length alone	Selected answer	R monitor	R (short only)	R (non-empty)
GPQA-Diamond	Qwen3-4B	0.97	0.516	0.518	0.505 ± 0.066	0.527 ± 0.086	0.505 ± 0.066
GPQA-Diamond	Qwen3-8B	0.55	0.638	0.544	0.631 ± 0.047	0.482 ± 0.151	0.631 ± 0.047
GPQA-Diamond	Qwen3-14B	0.99	0.512	0.545	0.558 ± 0.046	0.567 ± 0.052	0.558 ± 0.046
GPQA-Diamond	gpt-oss-20b	1.00	—	0.560	0.535 ± 0.137	0.535 ± 0.137	0.535 ± 0.137
GPQA-Diamond	gpt-oss-120b	1.00	—	0.538	0.542 ± 0.073	0.542 ± 0.073	0.542 ± 0.073
MMLU-Pro	Qwen3-4B	0.96	0.501	0.531	0.560 ± 0.014	0.551 ± 0.017	0.560 ± 0.014
MMLU-Pro	Qwen3-8B	0.66	0.530	0.524	0.652 ± 0.014	0.530 ± 0.016	0.652 ± 0.014
MMLU-Pro	Qwen3-14B	0.98	0.505	0.523	0.533 ± 0.004	0.528 ± 0.005	0.533 ± 0.004
MMLU-Pro	gpt-oss-20b	1.00	0.503	0.511	0.515 ± 0.009	0.515 ± 0.009	0.515 ± 0.009
MMLU-Pro	gpt-oss-120b	1.00	0.501	0.522	0.538 ± 0.006	0.538 ± 0.006	0.538 ± 0.006
Omni-MATH-2-Filtered	Qwen3-4B	0.06	0.617	0.593	0.807 ± 0.015	0.981 ± 0.013	0.782 ± 0.019
Omni-MATH-2-Filtered	Qwen3-8B	0.05	0.589	0.607	0.837 ± 0.027	0.928 ± 0.065	0.820 ± 0.028
Omni-MATH-2-Filtered	Qwen3-14B	0.13	0.529	0.606	0.779 ± 0.010	0.842 ± 0.040	0.752 ± 0.014
Omni-MATH-2-Filtered	gpt-oss-20b	0.45	0.628	0.567	0.682 ± 0.019	0.590 ± 0.039	0.682 ± 0.019
Omni-MATH-2-Filtered	gpt-oss-120b	0.60	0.611	0.568	0.671 ± 0.012	0.592 ± 0.034	0.671 ± 0.012

Appendix Table A72: **Restricting the analysis by trace availability and parser outcomes changes equal-weight mean AUROC by at most 0.010.** Blank-trace exclusion retains items whose whitespace-stripped reasoning trace is nonempty. The strict parsed-output restriction retains items with parser status "ok" and nonempty whitespace-stripped values for the reasoning trace, response, and extracted answer. AUROC is recomputed within each setting and data split after applying each restriction, averaged over five data splits within each model–benchmark setting, and then averaged equally over the 15 settings. Δ is restricted-sample minus full-sample AUROC; upper-block deltas are reported to four decimals because several are smaller than 0.001, and the per-setting block uses three decimals. The final column reports retained / full-sample held-out predictions across items and data splits for each access level. The lower block reports per-setting deltas for the two Omni-MATH-2-Filtered settings with concentrated empty traces.

Restriction	Access level	Full sample	Restricted sample	Δ	Scored predictions
Blank-trace exclusion	R	0.623	0.621	-0.0024	80,789 / 82,100
	$R{+}S$	0.779	0.776	-0.0030	80,789 / 82,100
	$R{+}S{+}C$	0.814	0.814	-0.0002	80,789 / 82,100
	$R{+}S{+}C{+}Z$	0.840	0.839	-0.0013	80,789 / 82,100
Strict parsed-output restriction	R	0.623	0.613	-0.0098	79,930 / 82,100
	$R{+}S$	0.779	0.772	-0.0070	79,930 / 82,100
	$R{+}S{+}C$	0.814	0.810	-0.0035	79,930 / 82,100
	$R{+}S{+}C{+}Z$	0.840	0.835	-0.0049	79,930 / 82,100

Affected-setting Δ (restricted-sample minus full-sample AUROC)

Restriction	Setting	R	$R{+}S$	$R{+}S{+}C$	$R{+}S{+}C{+}Z$
Blank-trace exclusion	Qwen3-8B / Omni-MATH-2-Filtered	-0.019	-0.029	-0.008	-0.011
	Qwen3-14B / Omni-MATH-2-Filtered	-0.007	-0.012	-0.003	-0.003
Strict parsed-output restriction	Qwen3-8B / Omni-MATH-2-Filtered	-0.059	-0.052	-0.029	-0.033
	Qwen3-14B / Omni-MATH-2-Filtered	-0.039	-0.035	-0.019	-0.019

Appendix Table A73: **Sparse models select both public feature sources.** Values are selection rates across the 15 model–benchmark settings and five random splits. Plain LASSO selects individual standardized features. Sparse group LASSO applies a penalty by feature source while retaining sparsity within each source. Both models include surface features, Cohere embeddings, and answer-probability features when available. Raw hidden states are evaluated separately in a sensitivity analysis using replay features.

Selection method	Feature source	R	$R+S$	$R+S+C$	$R+S+C+Z$
Plain LASSO	Surface	86%	96%	99%	97%
	Cohere	92%	100%	100%	100%
	Z_{prob}	—	—	—	76%
Sparse group LASSO	Surface	100%	97%	100%	100%
	Cohere	93%	92%	75%	84%
	Z_{prob}	—	—	—	79%

Appendix Table A74: **Stable sparse-feature examples are descriptive, not a complete explanation of the monitor.** The table lists human-readable surface features selected in at least three settings with mean selection rate at least 60%. The sign is the dominant coefficient direction among selected fits.

Analysis	Level	Feature	Settings	Selection	Sign
Sparse group LASSO	$R{+}S{+}C$	Mathematical-symbol density	15	100%	+
Sparse group LASSO	$R{+}S{+}C$	Sentence count	15	100%	−
Sparse group LASSO	$R{+}S{+}C$	Uncertainty-marker rate	15	100%	−
Sparse group LASSO	$R{+}S{+}C$	Word count	15	100%	−
Sparse group LASSO	$R{+}S{+}C{+}Z$	Answer-framing rate	15	100%	+
Sparse group LASSO	$R{+}S{+}C{+}Z$	Mathematical-symbol density	15	100%	+
Sparse group LASSO	$R{+}S{+}C{+}Z$	Uncertainty-marker rate	15	100%	−
Sparse group LASSO	$R{+}S{+}C{+}Z$	Word count	15	100%	−
Sparse group LASSO	$R{+}S{+}C$	Answer-framing rate	15	99%	+
Sparse group LASSO	$R{+}S{+}C$	Digit density	15	99%	+
Sparse group LASSO	$R{+}S{+}C{+}Z$	Digit density	15	99%	+
Sparse group LASSO	$R{+}S{+}C{+}Z$	Self-correction rate	15	99%	−
Sparse group LASSO	$R{+}S{+}C{+}Z$	Sentence count	15	99%	−
Sparse group LASSO	$R{+}S$	Digit density	15	96%	+
Sparse group LASSO	$R{+}S$	Word count	15	96%	−
Plain LASSO	$R{+}S{+}C$	Uncertainty-marker rate	15	96%	−
Sparse group LASSO	$R{+}S{+}C$	Self-correction rate	15	96%	−
Plain LASSO	$R{+}S{+}C{+}Z$	Uncertainty-marker rate	15	95%	−

Appendix Table A75: **Last90 contains denser and more discriminative marker cues than same-run self-summaries.** Entries are equal-weight means of setting-level statistics across 15 model–benchmark settings and cover $n{=}80{,}763$ paired runs after excluding 1,292 pairs with a zero-word Last90 display. No self-summary display has zero words. Δ is Last90 minus summary. Brackets are 95% descriptive setting-bootstrap intervals (10,000 resamples). Single-feature AUROC is oriented using the dominant coefficient direction in Appendix Table A74, except that average word length is positive by convention. Oriented values above 0.5 discriminate in the chosen direction. Last90 uses the whole trace when it is shorter than 90 whitespace-delimited words.

	Mean level			Oriented AUROC		
Feature	Summary	Last90	Δ [95% CI]	Summary	Last90	Δ [95% CI]
Word count	79.7	80.6	$+0.9$ [-9.3, $+10.8$]	0.693	0.659	-0.035 [-0.073, $+0.003$]
Sentence count	6.3	12.3	$+6.0$ [$+4.6$, $+7.5$]	0.488	0.382	-0.106 [-0.150, -0.064]
Average word length	4.74	4.25	-0.49 [-0.69, -0.28]	0.434	0.450	$+0.015$ [$+0.000$, $+0.031$]
Digit density	0.041	0.059	$+0.018$ [$+0.013$, $+0.024$]	0.634	0.612	-0.022 [-0.043, -0.003]
Math-character density	0.052	0.042	-0.010 [-0.020, -0.003]	0.576	0.619	$+0.044$ [$+0.025$, $+0.063$]
Uncertainty rate	0.003	0.013	$+0.009$ [$+0.008$, $+0.011$]	0.565	0.654	$+0.090$ [$+0.064$, $+0.118$]
Self-correction rate	0.007	0.019	$+0.012$ [$+0.010$, $+0.014$]	0.521	0.589	$+0.068$ [$+0.048$, $+0.089$]
Answer-marker rate	0.015	0.066	$+0.051$ [$+0.046$, $+0.056$]	0.625	0.696	$+0.072$ [$+0.049$, $+0.095$]