# Measuring Black-Box Confidence via Reasoning Trajectories: Geometry, Coverage, and Verbalization

Marc Boubnovski Martell[1] Josefa Lia Stoisser[1] Kaspar Märtens[1] Jialin Yu[2,3]
Robert Kitchen[1] Philip Torr[2] Jesper Ferkinghoff-Borg[1]
[1]Novo Nordisk [2]University of Oxford [3]University of Bath

## Abstract

Reliable confidence estimation gates safe deployment of chain-of-thought (CoT) reasoning through text-only APIs, yet the dominant black-box baseline, self-consistency over $K$ samples, is linearly expensive and ignores the geometry of the generated text trajectory. We introduce a black-box trajectory-confidence score, *Geometry* ($G$), that embeds a CoT as a sliding-window trajectory and measures its convergence toward external candidate anchors with a one-parameter softmax, requiring no access to token probabilities, hidden states, or generator weights at inference. On six $(\text{benchmark}, \text{reasoner})$ settings over MedQA-USMLE, GPQA Diamond, and MMLU-Pro $\times$ {Gemini 3.1 Pro, Claude Sonnet 4.6}, fusing $G$ with coverage and verbalized-confidence channels at $K{=}4$ improves over self-consistency at $K{=}8$ in $6/6$ settings (median AUC $0.78$ vs. $0.71$, $\Delta\text{AUC} = +0.075$); a fixed-pick control ($+0.060$) and an E5 cross-embedder replication rule out answer-switching and single-vendor artifacts. Because this full configuration routes coverage through a reference-conditioned judge, we report it explicitly as *reference-assisted*, and separate it from the verifier-free operating points $G$ and $G{+}V$. Controlled black-box interventions support a pairing-specific, position-dependent reading of $G$: replacing a trace with another question's trajectory while holding candidate anchors fixed collapses median AUROC from $0.744$ to $0.425$–$0.500$ across five swap constructions, and body/final-answer substitutions show that pre-answer geometry is trajectory-specific whereas terminal geometry is answer-string driven. Three increasingly unscaffolded regimes decompose black-box confidence into a judge-mediated Coverage channel ($C$), within-trace Geometry ($G$), and a conditional Verbalization channel ($V$); across 18 benchmark$\times$reasoner$\times$proposer settings, likelihood-ratio tests find incremental predictive value conditional on the other channels for $C$ in $18/18$ and $G$ in $16/18$, while $V$ contributes in only $6/18$. Together, these results show that black-box CoT confidence can be read from a trace's geometric convergence in embedding space without token-level probability access, and that the appropriate configuration depends on whether a trusted reference or verifier is available.

## 1 Introduction

Chain-of-thought (CoT) reasoning Wei et al. [2022] is often deployed through text-only APIs Stoisser et al. [2025a], where logits, hidden states, and retrieval support are unavailable and confidence must be estimated from the generated text alone Kuhn et al. [2023], Lin et al. [2023], Pedapati et al. [2024]. The dominant black-box baseline is self-consistency: sample many CoTs and use vote concentration as confidence Wang et al. [2022b], Martell et al. [2026]. It is effective, but linearly expensive in $K$, blind to the geometry of the generated text trajectory, and produces an inherently

*discrete* confidence, a $k/K$ ratio over a finite ballot, which loses resolution at small $K$ and cannot rank ties Martell et al. [2025]. We study a different signal: the trajectory of the CoT itself. By embedding sliding windows of the trace and scoring their convergence toward external candidate anchors with a *one-parameter softmax calibration model* (a single inverse-temperature slope $\beta$ fit by maximum likelihood), we read out a *continuous numerical* black-box confidence score from text alone, one logit-shaped probability per option per question, with the same sign and significance of $\hat{\beta}$ under three embedders (`gemini-embedding-001`, OpenAI `text-embedding-3-large`, E5 Wang et al. [2022a]). Throughout, "trajectory geometry" refers to the geometry of the *generated text* in an external embedding space; we do not observe the generator's internal representations.

Multiple-choice benchmarks provide the cleanest controlled setting because they supply both candidate anchors and a gold answer. We use three regimes: visible-option multiple-choice (MCQ; multiple-choice question answering, MCQA, when referring to the task) as a sanity check, blinded option anchors as the main single-channel configuration, and an open-ended proposer/judge regime where even the candidate set is generated rather than given. In the blinded regime the fitted slope of the one-parameter softmax (formally $\hat{\beta}$ in Section 3) is strongly negative in every benchmark–model combination, the geometry-based composite reaches AUC $0.752$–$0.826$ on MedQA and $0.701$–$0.727$ on GPQA, and the window-position sweep gives a behavioural, text-level diagnostic of *where* in the trace the signal concentrates.

But single-trace geometry has a structural ceiling: if no sampled trace ever reaches the correct hypothesis, no score defined on that trace can recover it. Moving to the open-ended regime separates this coverage failure from within-trace commitment failure. This yields three black-box channels with *different access requirements*: Coverage ($C$), Geometry ($G$), and Verbalization ($V$). $G$ is candidate-conditional and requires no per-question reference at inference; $V$ is a reasoner self-report; $C$, as implemented here, requires a per-question reference answer or another trusted verification signal. We therefore report $G$ as the paper's primary novel trace-derived signal and the full $C{+}G{+}V$ fusion as a *reference-assisted* headline that improves on matched self-consistency at comparable end-to-end API cost (accounting for proposer, judge, verbalization, and embedding calls; Appendix E.4), and we treat verbal-only, $G$, $G{+}V$, and $C{+}G{+}V$ as distinct operating points rather than one mandatory pipeline (Section 5.3).

**Contributions.**

1. **A black-box trajectory-confidence score.** A one-parameter sliding-window cosine softmax mapping $(\text{text}, \text{embedder}, \text{anchors}) \mapsto$ confidence, replicating across three embedders, three reasoners, and three benchmarks. $G$ is trace-derived and candidate-conditional: it requires candidate anchors, but needs no per-question reference or verifier at inference.
2. **Pairing-specific and position-dependent signal.** Holding anchors fixed and swapping trajectories collapses median AUROC from $0.744$ to $0.425$–$0.500$ across five controls. Body/final-answer substitutions show pre-answer body geometry is trajectory-specific, whereas terminal geometry is answer-string driven (selecting appended counterfactual wrong answers in $80.4\%$ of cases; Section 4.4).
3. **Three-channel decomposition and asymmetric $V$.** In open-ended MCQA, black-box confidence decomposes into judge-mediated Coverage ($C$), within-trace Geometry ($G$), and conditional Verbalization ($V$), with incremental predictive value in $18/18$ ($C$), $16/18$ ($G$), and $6/18$ ($V$) settings. Fused $C{+}G{+}V$ at $K{=}4$ improves over SC@8 in all six closed-set settings at comparable API cost (median $\Delta\text{AUC} = +0.075$, Stouffer $z = 5.67$), confirmed by a fixed-pick control ($+0.060$).

## 2 Related Work

Methods for estimating LLM confidence differ along three axes: *access* (white-box logits/activations vs. black-box text), the *object* scored (final answers vs. intermediate traces), and the *reference* used (self-referential agreement vs. external anchors). A *candidate anchor set* lists hypotheses against which a trace is scored (supplied by the benchmark, an external system, or a proposer model) without indicating correctness; a *reference signal* identifies the correct candidate. Our Geometry channel requires only candidate anchors; Coverage additionally requires a reference or trusted verifier.

White-box methods rely on token probabilities, post-hoc calibration, hidden-state probes, or representation geometry Shorinwa et al. [2025], Jiang et al. [2020], Kadavath et al. [2022], Azaria and Mitchell [2023], Li et al. [2023], Belrose et al. [2023], Ali et al. [2025], which commercial APIs typically withhold. Black-box alternatives evaluate verbalized confidence Lin et al. [2022], Tian et al. [2023], Xiong et al. [2023], Yoon et al. [2025], self-consistency Wang et al. [2022b], semantic entropy Farquhar et al. [2024], or answer-set agreement Manakul et al. [2023], Agrawal et al. [2024], but score final outputs or self-reports rather than intermediate reasoning geometry Ke et al. [2025]. We evaluate two verbal baselines: a standard 0–100 prompt and a post-hoc protocol adapted from Yoon et al. [2025] (Appendix H, Section 4.6).

Closest to our work are geometric analyses of generations: Lin et al. [2023], Phillips et al. [2025] score dispersion in clouds of completed responses, while Jiang et al. [2026] analyzes kinematics of reasoning traces; neither scores against external answer anchors. Process-reward models require step supervision Lightman et al. [2023]; CoT-faithfulness studies whether surface text reflects underlying computation Turpin et al. [2023], Lanham et al. [2023]. Downstream, confidence gates tool use and synthetic data curation Stoisser et al. [2025b,c]. Concurrently, Sun et al. [2026] find late-stage divergence in hidden states; our work is complementary, operating black-box on text trajectories with external anchors. We evaluate under selective classification Cole et al. [2023], Wen et al. [2025], matching real-world abstention needs.

## 3 Isolating the Geometric Channel

We first isolate **Geometry** ($G$), the within-trace spatial commitment of a CoT trajectory toward an external answer anchor. On closed-set MedQA, GPQA, and MMLU-Pro, $G$ alone reaches single-feature selective AUC 0.71–0.82 with no proposer, judge, or self-report. **Coverage** ($C$) and **Verbalization** ($V$) become necessary only in the open-ended regime of Section 5, where a trace can fail by missing the gold ($G$ cannot recover) or by committing to the wrong in-set hypothesis ($C$ cannot recover). This section formalizes $G$ alone.

We isolate the trajectory-geometry signal in a *blinded multiple-choice protocol*: the generator reasons about a closed-set question without seeing the option list, and we score its trace against the dataset options externally. This forces the CoT-to-option mapping to be performed geometrically rather than by the generator itself. A visual schematic of this scoring pipeline is given in Appendix A.1 (Figure 1).

### 3.1 Setup and Trajectory Construction

A multiple-choice question $q$ has $J{=}4$ options $\{a_1, \ldots, a_J\}$ with gold index $y^\star$. The gold index is used for *training and evaluation* of the calibration parameters and for reporting AUC; computing $G$ at inference requires candidate anchors but not a per-question gold or gold-equivalence signal. This is the property that separates $G$ from the Coverage channel of Section 5, whose judge-mediated features are computed against a reference answer; we make the resulting deployment distinction explicit in Section 5.3. We prompt the generator with $q$ alone, withholding the option list, and sample $K{=}4$ free-form chains of thought $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(K)}$ at $T{=}0.7$. The target is (i) a pick $\hat{y} \in \{1, \ldots, J\}$ and (ii) a confidence $c \in [0, 1]$ for $\{\hat{y} = y^\star\}$. Withholding options at generation time forces the CoT-to-option map to be performed externally and geometrically rather than by the generator itself.

Each CoT is embedded as a trajectory of overlapping sliding windows under embedding model $\phi$ : text $\to \mathbb{R}^p$ (`gemini-embedding-001`, $p{=}3072$). With window size $W{=}30$ and stride $S{=}15$ words: $\mathcal{T}^{(k)} = \big(\mathbf{z}_0^{(k)}, \mathbf{z}_1^{(k)}, \ldots, \mathbf{z}_{M_k}^{(k)}\big)$, where $\mathbf{z}_0^{(k)} \equiv \phi(q)$ and $\mathbf{z}_m^{(k)} = \phi(w_{(m-1)S+1:\,(m-1)S+W}^{(k)})$ for $m \geq 1$, with $M_k$ total windows. Option anchors are the embeddings of benchmark option texts, $\mathbf{y}_j = \phi(a_j)$, fixed across samples. The trajectory-to-anchor distance is the cosine distance from the *penultimate* window $\mathbf{z}_{M_k-1}^{(k)}$ to each anchor, averaged over the $K$ samples:

$$D_j^{(k)} \;=\; 1 - \frac{\langle \mathbf{z}_{M_k-1}^{(k)},\, \mathbf{y}_j \rangle}{\|\mathbf{z}_{M_k-1}^{(k)}\|\, \|\mathbf{y}_j\|}, \qquad D_j \;=\; D_j(\mathcal{T}) \;=\; \frac{1}{K} \sum_{k=1}^{K} D_j^{(k)}. \tag{1}$$

Scoring the penultimate window guards against terminal answer-restatement (Section 4.4, Appendix C). Euclidean and PCA variants behave similarly (Section 4.5).

### 3.2 Scoring Rule and Composite Confidence

**Geometric pick and softmax probability.** The pick is the closest anchor in mean cosine distance, $\hat{y} = \arg\min_j D_j$, with correctness label $\ell = \mathbb{1}[\hat{y} = y^\star]$ serving as the target of every selective-prediction AUC reported in this paper. Conditional on the option set, we score correctness with a one-parameter softmax over option distances,

$$\textsc{Geo}(q) \;=\; \frac{\exp(\hat{\beta}_{\text{all}}\, D_{\hat{y}})}{\sum_{j'} \exp(\hat{\beta}_{\text{all}}\, D_{j'})}, \qquad \hat{\beta}_{\text{all}} \;=\; \arg\max_{\beta} \sum_{q'} \log \frac{\exp(\beta\, D_{y^\star_{q'}})}{\sum_{j'} \exp(\beta\, D_{j',q'})}, \tag{2}$$

fit on the training fold. A strongly negative $\hat{\beta}_{\text{all}}$ means closer anchors are more likely correct; we verify this empirically in Section 4.2, and Table 6 reports $\hat{\beta}_{\text{all}}$ per (dataset, model). The softmax probability $\textsc{Geo}(q)$ is our single geometric feature.

**One posterior, three readouts.** GEO, the null-standardized top-two margin $\tilde{D} = (D_{(2)} - D_{(1)} - \mu_q)/\sigma_q$ (with $\mu_q, \sigma_q$ estimated from a cross-question null pool), and the binary correctness probability $P(\hat{y} = y^\star \mid \mathcal{T})$ are three projections of one Bayesian posterior with uniform priors (Appendix I). The binary projection is the Fermi–Dirac form:

$$P(\hat{y} = y^\star \mid \mathcal{T}) = \sigma(\hat{\beta}_{\text{bin}}\, \tilde{D}(\mathcal{T}) - \alpha), \qquad \sigma(t) = 1/(1 + e^{-t}), \tag{3}$$

where the all-option and binary fits satisfy $\hat{\beta}_{\text{all}} = -\hat{\beta}_{\text{bin}}/\bar{\sigma}_q$ with $\bar{\sigma}_q$ the mean null scale (Eq. 4, verified in Appendix Table 23).

$$\hat{\beta}_{\text{all}} = -\hat{\beta}_{\text{bin}}/\bar{\sigma}_q. \tag{4}$$

**Composite confidence.** We fuse GEO with four auxiliary features on the same $K{=}4$ samples: STR (string consensus of condensed answers), SEM (mean pairwise cosine among answer embeddings), VOL (sample/window-averaged consecutive-window step distance), and ENS (cross-model modal agreement). Condensed answers use a zero-shot prompt (Appendix H); formulas are in Appendix G. The composite is the predicted probability from a logistic regression on standardized (GEO, STR, SEM, VOL, ENS) with balanced class weights. The **No-Geo** baseline refits on the other four features, isolating $\Delta(\textsc{Geo})$.

**Protocol.** Parameters ($\hat{\beta}_{\text{all}}$, standardizer, logistic weights) are fit on training folds of a stratified 5-fold split and evaluated frozen on held-out folds; we report mean test AUC across folds. Auxiliary features drop questions where condensation fails, yielding retained $N$ of $1272/1271$ (MedQA) and $117/187$ (GPQA) for Gemini/Sonnet. Appendix C gives the corrected natural-window audit.

## 4 Validating the Geometric Measurement

### 4.1 Experimental Setup

We evaluate on MedQA-USMLE Jin et al. [2021] (1,273 questions, 4 options; retained $N{=}1272/1271$ for Gemini/Sonnet after parse filtering) and GPQA Diamond Rein et al. [2023] (198 questions; retained $N{=}117/187$). The generator sees only the question; we sample $K{=}4$ open-ended CoTs at $T{=}0.7$, condense each to a predicted answer, and evaluate via stratified 5-fold CV whether the geometric argmin-pick is correct. Primary reasoners are **Gemini 3.1 Pro** and **Claude Sonnet 4.6**, supplemented by a 200-question MMLU-Pro pilot.

### 4.2 Result 1: A One-Parameter Softmax Tracks Correctness

Trajectory endpoint distance strongly predicts correctness in the blinded option-anchor protocol: across all four MedQA/GPQA combinations, the fitted softmax slope $\hat{\beta}$ is strongly negative ($z = -5.48$ to $-25.39$), with single-feature GEO AUC reaching 0.71–0.82 (Appendix Table 6; third-benchmark and open-weight Llama-3.3 replications in Appendix Tables 8, 14). Falsification tests confirm that bootstrap CIs of $|\hat{\beta}_{\text{all}}|$, $\hat{\beta}_{\text{bin}}$, and $\hat{\beta}_{\text{bin}}/\bar{\sigma}_q$ overlap pairwise in $6/6$ combinations (Appendix Table 23), verifying unified-posterior consistency.

Table 1: **Selective-prediction AUC in the blinded option-anchor protocol.** Geo is the fitted softmax probability of the picked option; Str/Sem/Vol/Ens are auxiliary agreement features; Composite is a logistic fusion and No-Geo removes Geo. Geo is en-par or better than Composite in all four settings.

| Dataset | Model | Geo | Str | Sem | Vol | Ens | Composite | No-Geo | $\Delta$(Geo) |
|---|---|---|---|---|---|---|---|---|---|
| MedQA | Gemini 3.1 Pro | $\mathbf{0.821}$ | $0.512$ | $0.562$ | $0.558$ | $0.506$ | $\mathbf{0.826}$ | $0.577$ | $+0.249$ |
| MedQA | Claude Sonnet 4.6 | $\mathbf{0.753}$ | $0.536$ | $0.549$ | $0.490$ | $0.508$ | $\mathbf{0.752}$ | $0.519$ | $+0.234$ |
| GPQA | Gemini 3.1 Pro | $\mathbf{0.710}$ | $0.579$ | $0.530$ | $0.593$ | $0.552$ | $0.701$ | $0.560$ | $+0.141$ |
| GPQA | Claude Sonnet 4.6 | $\mathbf{0.750}$ | $0.628$ | $0.466$ | $0.496$ | $0.538$ | $0.727$ | $0.605$ | $+0.122$ |

### 4.3 Result 2: A Single Confidence Signal Across Domains and Models

Does trajectory geometry contribute signal beyond answer-set agreement? We fuse GEO with four consistency and ensemble covariates defined in Appendix G, and compare against a No-Geo ablation that refits the same logistic without the geometric feature. Table 1 shows that the composite reaches AUC $0.701$–$0.826$ across all four benchmark–model combinations, while removing Geo costs $0.12$–$0.25$ AUC. Geo is strongest on MedQA and remains competitive with, or stronger than, the full composite on GPQA.

**Single-channel ($G$-only) comparison to $K$-sample self-consistency.** As a sanity check before the full three-channel decomposition of Section 5, single-channel $G$ already exceeds budget-matched self-consistency Wang et al. [2022b] (SC@$K$, the plurality vote share over $K$ CoTs after mapping sampled answers into the benchmark's answer space) in $9/10$ benchmark–model–condition combinations (Appendix Tables 9 and 10). Semantic entropy is treated as a separate baseline rather than as the SC voting rule: DeBERTa-MNLI semantic entropy Kuhn et al. [2023], Farquhar et al. [2024] on a MedQA/Gemini pilot reaches AUC $0.590$ versus $0.821$ for the geometric composite. The fully matched $C{+}G{+}V$ vs. SC@8 comparison is the headline of Section 5.4.

### 4.4 Result 3: Position and Pairing Diagnostics of the Geometric Signal

Does the geometric signal depend on the specific pairing between a trace and its candidate anchors, and where does it concentrate? We test this via window sweeps and two black-box interventions.

**Corrected window-position analysis.** An audit revealed that the positive GPQA coefficient previously reported at the terminal window $k{=}{-}1$ was computed from an *appended bare answer label* (e.g., "The answer is (B)") rather than natural CoT text. On natural CoT windows, the terminal coefficient is strongly negative ($\hat{\beta}_{\text{all}} = -42.6$ Gemini, $-28.8$ Claude), and the penultimate window is more negative still ($-48.2$ and $-39.5$). Replacing the bare label with full option text restores negative coefficients ($-47.4$, $-36.5$), while appending a question-invariant constant label reproduces the positive artifact ($\approx +20$). We thus withdraw the terminal answer-commitment reversal claim. The penultimate natural window remains empirically optimal and avoids endpoint contamination (Appendix C).

**Trajectory swaps.** Holding candidate anchors fixed, replacing trajectories with those from other questions collapses median AUROC from $0.744$ to $0.475$ (random), $0.475$ (same-pick), $0.491$ (different-pick), $0.500$ (nearest-trajectory), and $0.425$ (nearest-anchor; Table 2; $N{=}196$–$200$/cell, $2{,}000$ bootstrap resamples). The collapse under nearest-trajectory and nearest-anchor replacements confirms that neither topical similarity nor selected-option agreement preserves the signal once question-specific trajectory–anchor pairing is severed.

**Body/final-answer intervention.** We next separate the pre-answer "body" of a trace from its final answer sentence. Original pre-answer bodies yield median candidate-pick accuracy $0.624$; replacing each body with its *nearest-neighbour* body from another question, the hardest available substitution, drops this to $0.372$ and produces a negative median gold margin ($-0.007$). Appending the original question's final answer sentence to a swapped body recovers original-question pick accuracy to $0.631$, essentially the unswapped baseline, whereas appending the swapped trajectory's own final answer

Table 2: **Trajectory-swap control.** Each question's candidate anchors are held fixed and its trajectory is replaced by another question's; we rescore the original pick. AUROC is of the softmax probability at the original pick against the original correctness label, with 95% percentile bootstrap CIs over 2,000 question resamples. Median is across the four cells.

| Cell | $N$ | **Correctly paired** | Random | Same-pick | Different-pick | Nearest-traj. | Nearest-anchor |
|---|---|---|---|---|---|---|---|
| GPQA / Gemini | 196 | 0.641 [0.560, 0.716] | 0.482 | 0.473 | 0.445 | 0.456 | 0.515 |
| GPQA / Claude | 197 | 0.639 [0.561, 0.714] | 0.469 | 0.395 | 0.477 | 0.543 | 0.459 |
| MedQA / Gemini | 200 | 0.924 [0.882, 0.959] | 0.480 | 0.507 | 0.505 | 0.559 | 0.346 |
| MedQA / Claude | 199 | 0.847 [0.790, 0.898] | 0.347 | 0.477 | 0.510 | 0.441 | 0.391 |
| **Median** | — | **0.744** | 0.475 | 0.475 | 0.491 | 0.500 | 0.425 |

Table 3: **Strong verbal and privileged-likelihood baselines on GPQA Diamond with GPT-4.1.** Cross-validated AUROC for predicting correctness of model's extracted answer on aligned same-trace subset ($N$=194, accuracy 0.613) on which Geometry was recomputed, alongside unaligned full logprob run ($N$=198, acc 0.601).

| **Signal** | **Access** | $N$ | **AUROC** |
|---|---|---|---|
| Verbal confidence, adapted Yoon-style prompt | black-box | 194 | 0.648 |
| Geometry, top-two margin | black-box | 194 | 0.629 |
| $G{+}V$ (calibrated probability + verbal) | black-box | 194 | **0.698** |
| $G{+}V$ (margin + verbal) | black-box | 194 | **0.705** |
| Final answer-token likelihood | privileged | 198 | 0.477 |
| Full-response mean log-probability | privileged | 194 | 0.663 |
| Exact 90%-prefix intervention | privileged | 194 | 0.594 / 0.603 |
| $G$ + prefix intervention | mixed | 194 | 0.649 / 0.657 |

leaves it at 0.333. Conversely, appending a counterfactual *wrong* answer to the original body causes terminal-window geometry to select the appended wrong answer in 80.4% of cases. Per-cell values and bootstrap CIs are given in Appendix Table 13. The two regimes are thus dissociated: pre-answer body geometry carries trajectory-specific information, while terminal-window geometry is dominated by the literal answer string. This is the concrete reason the paper scores the penultimate window.

**Interpretation.** These interventions provide controlled black-box behavioural evidence for a pairing-specific and position-dependent text-level signal. They do not identify the generator's hidden computation or establish a causal internal mechanism, and we make no such claim.

### 4.5 Ablations

Ablations confirm stability: metric/PCA variations preserve $\hat{\beta}$ significance; $W \in \{20, 30, 40\}$ is stable; swapping to OpenAI `text-embedding-3-large` lifts the composite over SC in $7/7$ open-ended settings; and trace-length residualization shifts AUC by $\leq +0.008$. Anchor design is critical: trajectory-derived anchors collapse to near chance, while external dataset/rationale anchors succeed (Appendix B).

### 4.6 Result 4: Strong Verbal and Privileged-Likelihood Baselines

We evaluate against strong verbal self-reports and privileged token probabilities on GPT-4.1 (GPQA Diamond, same-trace comparison; Table 3). On aligned traces ($N$=194, acc 0.613), an adapted Yoon-style verbal prompt reaches AUROC 0.648, slightly above standalone Geometry margin (0.629). Fusing the two yields the best black-box performance ($G{+}V = 0.705$, $\Delta = +0.057$, 95% CI $[-0.009, +0.132]$), aided by low cross-channel correlation ($r$=0.15). Verbal confidence is poorly calibrated in level (mean conf 0.930, ECE = 0.320), serving primarily as a ranking signal. On the privileged side, final-token likelihood is near chance (0.477), whereas full-response mean log-probability reaches 0.663 ($N$=194). Geometry is thus competitive with, rather than dominant over, strong verbal and likelihood baselines, providing an effective black-box signal when token probabilities are inaccessible.

We draw a deliberately restricted conclusion. Geometry does *not* uniformly dominate either verbalized confidence or likelihood-based scores. It is a competitive black-box signal that carries partially complementary information without requiring token-level probability access, and the practically relevant comparison is therefore between access regimes rather than a single ranking of estimators.

# 5 Three Channels of Black-Box Confidence

Section 4 established that $G$ is a valid black-box measurement: the geometric score tracks correctness in every benchmark–model setting and beats budget-matched self-consistency in $9/10$ single-channel settings. This section asks the orthogonal deployment question: *is one channel sufficient?* We argue (and quantify) that it is not, that the residual failure decomposes into a coverage failure and a commitment failure, and that a three-channel fusion at $K{=}4$ improves on SC@8 in $6/6$ closed-set settings. The three channels are not, however, three views of the same reasoner-side quantity. $G$ and $V$ are reasoner-side signals, derived respectively from the generated trace and from the generator's self-report. $C$ is an external verification signal: it measures whether a reference-conditioned judge can find the correct answer in a proposer's hypothesis pool. $G$ alone cannot detect that a candidate space is inadequate, because it scores commitment relative to whatever candidates it is given. In reference-assisted settings $C$ addresses exactly that failure; without a verifier, $C$ is unavailable and the failure remains uncovered.

## 5.1 Why a Second Channel Is Necessary

The geometric score measures how strongly a CoT trace concentrates near one option region in embedding space; it is invariant to whether that region happens to contain the gold option. To make this concrete, we stratify the closed-set blinded protocol into three buckets: (i) CORRECT (the picked option is gold); (ii) WRONG-IN-SET (gold was proposed by some sample but not picked); and (iii) NO-HYPOTHESIS-CORRECT (gold was not proposed by any sample). Bucket (ii) and bucket (iii) are both "wrong" from the user's standpoint, so a calibrated confidence score should rank them equivalently low. It does not. Across all six closed-set $(\text{benchmark}, \text{model})$ combinations, every embedding-based and vote-based score we evaluate, GEO, the Bayesian top-2 projection, self-consistency vote share, normalized cluster entropy, and an anchor-cohesion control ($-\bar{\rho}$, with the cloud-cohesion $\rho_q$ defined in Appendix I.1), ranks bucket (iii) *above* bucket (ii), with AUCs in the $0.06$–$0.44$ range (i.e. systematically anti-correlated with correctness; Appendix I.1). The single score family that escapes this inversion is verbalized confidence (AUC $0.45$–$0.62$), which the same model produces conditional on its own CoT and which depends on no embedding geometry. The asymmetry is not numerical noise; it is a structural fingerprint of the channel.

The mechanism is direct. Under the parameterization of Section 3.2, the geometric channel is fully summarized by per-option distances; bucket-(iii) traces simply have small $D$ to their (in-set, but wrong) committed option, indistinguishable from bucket-(i) traces by construction. The calibration knobs internal to the channel, increasing $K$, swapping the anchor variant (condensed-answer vs. reasoning-tail vs. dataset-option, App. B), rescaling the softmax temperature $\beta$, and reweighting via a rejection prior over options, all preserve the bucket ordering or trivialize it (Appendix I.1). Reliable abstention therefore requires a second channel whose evidence does not flow through option-distance likelihoods, and verbalization is the only black-box source we have that meets that constraint.

## 5.2 Setup: Proposer, Judge, and the Three Channels

We replace the closed-set option list with a hypothesis pool generated by an independent *proposer* LLM and delegate gold-equivalence to a third *judge* LLM. The three agent roles are: a reasoner (Gemini 3.1 Pro, Claude Sonnet 4.6, or Llama 3.3 70B; $K{=}5$ CoT traces), a proposer (Gemini 3.1 Pro or Claude Sonnet 4.6; $K_P{=}3$ temperature-varied hypothesis pools of $J{=}4$ hypotheses each), and a fixed judge (Azure-hosted `openai_gpt5_mini`, publication date 2025-08-07, called once per pool under strict matching, with inputs and outputs archived for re-evaluation). What is new here is not proposer–judge plumbing per se but treating proposer-pool reachability of the gold answer as a first-class *confidence channel*, distinct from within-trace geometric concentration. This separates two failure modes that closed-set MCQA conflates, the model never raised gold ($C$) versus the model raised gold but did not commit ($G$), and assigns them to different model families so that no single vendor's idiosyncrasies drive both signals. Figure 2 (Appendix A.2) summarizes the

resulting information flow across the three agent roles and three feature channels. The bucket-(iii) failure of Section 5.1 now splits cleanly into two sub-cases that closed-set MCQA cannot separate: `absent_gold` (proposer never raised the gold) is the coverage failure; `no_path_in_set` (gold raised but $v_{\text{gold}}=0$ in the reasoner's votes) is the commitment failure. We define three confidence channels, each tied to one source of evidence:

- **Coverage** ($C$). Features are $\bar{c}_q$, $K_{P,q}^{\text{valid}}$, and the entropy $H_q$ of judge labels across valid pools, all defined in Eq. 5 below. *No reasoner trace is consumed.*
- **Geometric** ($G$). Trajectory-anchor geometry as in Section 3, with proposer-hypothesis anchors in place of dataset options. Features: top-2 cosine margin, mean cosine to top-1 hypothesis, vote concentration, vote entropy across $K=5$ traces.
- **Verbalized** ($V$). The same generator is shown its own CoT (no option list, no proposer pool) and asked for an integer 0–100 confidence in its committed answer. Features: mean, std, min over $K$ scores.

**Definition of Coverage features.** For question $q$ and pool $p \in \{1, \ldots, K_P\}$, the proposer emits $J=4$ hypotheses $\{a_{q,p,j}\}$. The judge returns $\ell_{q,p} \in \{1, \ldots, J\} \cup \{\text{NONE}\}$, indicating which hypothesis matches the reference answer. Over valid pools $\mathcal{P}_q^{\text{valid}}$ ($K_{P,q}^{\text{valid}} = |\mathcal{P}_q^{\text{valid}}|$), the pool-level indicator, question coverage score, and label entropy are:

$$c_{q,p} \;=\; \mathbb{1}[\ell_{q,p} \neq \text{NONE}], \qquad \bar{c}_q \;=\; \frac{1}{K_{P,q}^{\text{valid}}} \sum_{p \in \mathcal{P}_q^{\text{valid}}} c_{q,p}, \qquad H_q \;=\; -\sum_{v \in \mathcal{V}_q} \hat{p}_q(v) \log \hat{p}_q(v), \tag{5}$$

where $\hat{p}_q(v)$ is label frequency. With $K_P=3$, $\bar{c}_q \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ is a coarse four-level variable and our strongest single feature. Crucially, the judge is conditioned on the reference answer, making $\bar{c}_q$ reference-conditioned (Section 5.3). The open-ended panel uses $K=5$ reasoner samples; the closed-set protocol uses $K=4$.

**Combining the channels.** Let $\mathbf{x}_h(q)$ denote the feature vector defined above for channel $h \in \{C, G, V\}$. We use two-stage logistic stacking:

$$\begin{aligned} s_h(q) &= \sigma\big(b_h + \mathbf{w}_h^\top \operatorname{std}_h[\mathbf{x}_h(q)]\big), \qquad h \in \{C, G, V\}, \\ \widehat{p}_{CGV}(q) &= \sigma\left(b_F + \sum_{h \in \{C,G,V\}} w_{F,h} \operatorname{std}_{F,h}[s_h(q)]\right), \qquad \sigma(t) = (1 + e^{-t})^{-1}. \end{aligned} \tag{6}$$

Here std uses componentwise training-fold means and standard deviations (plus $10^{-9}$). Both stages learn $\ell_2$-regularized logistic weights against binary correctness. First-stage out-of-fold scores enter a separately split five-fold fusion; this is not fully nested, since first-stage models are not retrained within fusion folds. Here $C$ is reference-conditioned coverage, not vote share.

### 5.3 Deployment Requirements of the Three Channels

The three channels have fundamentally distinct access requirements:

- **Coverage ($C$): Reference-assisted.** As implemented here, $C$ requires a reference answer or trusted verifier at inference to construct the judge label; in fully open-ended settings without verification, $C$ is unavailable. In structured agentic workflows, an external verification signal could in principle be supplied by executable tests, tool outcomes, or expert review, though this is not evaluated here.
- **Geometry ($G$): Verifier-free, candidate-conditional.** $G$ requires no reference or verifier at inference, but depends on candidate anchors (benchmark options or proposer hypotheses). $G$ measures commitment relative to the candidate set provided and cannot detect if that set lacks a viable answer (Appendix B). Candidate adequacy is thus an external prerequisite.
- **Verbalization ($V$): Reasoner-intrinsic.** Requires only query access for post-hoc self-report.

Consequently, the verifier-free configurations are $G$ and $G{+}V$. We treat $V$, $G$, $G{+}V$, and reference-assisted $C{+}G{+}V$ as distinct operating points along an access–cost–performance frontier (Appendix Table 19).

Table 4: **Three-channel selective-prediction AUC in open-ended MCQA** (median over two proposers per pair). **Bold**: best single channel. Fused $C{+}G{+}V$ (5-fold CV logistic) beats the best single channel in 8/9 rows. Full 18-cell panel in Appendix Table 15.

| Bench | Reasoner | C | G | V | C+G+V |
|---|---|---|---|---|---|
| GPQA | Gemini | **0.82** | 0.74 | 0.69 | 0.87 |
| GPQA | Claude | **0.75** | 0.74 | 0.71 | 0.85 |
| GPQA | Llama | **0.68** | 0.64 | 0.56 | 0.75 |
| MedQA | Gemini | **0.75** | 0.68 | 0.59 | 0.82 |
| MedQA | Claude | 0.73 | 0.66 | **0.76** | 0.86 |
| MedQA | Llama | **0.73** | 0.64 | 0.58 | 0.78 |
| MMLU-Pro | Gemini | **0.72** | 0.70 | 0.57 | 0.75 |
| MMLU-Pro | Claude | 0.64 | **0.67** | 0.65 | 0.70 |
| MMLU-Pro | Llama | **0.67** | 0.61 | 0.53 | 0.67 |
| Median across all 9 pairs | | 0.73 | 0.67 | 0.59 | **0.78** |

### 5.4 Decomposition Results and Headline Gain

Across the six closed-set (benchmark, model) settings spanning MedQA, GPQA Diamond, and MMLU-Pro $\times$ {Gemini 3.1 Pro, Claude Sonnet 4.6}, reference-assisted $C{+}G{+}V$ at $K{=}4$ outperforms SC@8 in AUC in $6/6$ settings (median $\Delta\text{AUC} = +0.075$, mean $+0.077$, pooled Stouffer $z = 5.67$, $p = 7.1 \times 10^{-9}$;[1] $4/6$ bootstrap CIs exclude zero). Holding the chosen answer fixed at SC's plurality pick on the 18-setting panel of Table 4, the geometric margin at SC's pick beats SC's vote share by median $+0.060$ AUC under 5-fold CV (SC and Geo agree on the pick in $94\%$ of questions), proving the gain reflects calibration rather than pick-switching. Same-trace replication with E5-large-v2 leaves $C{+}G{+}V$ above SC@8 in $6/6$ settings (pooled $z = 3.14$, $p = 8.5 \times 10^{-4}$, Appendix E.2), ruling out single-vendor artifacts.

Table 4 reports proposer-median AUC across proposers (full 18-cell panel in Appendix Table 15). Coverage empirically acts like a question-level prior while remaining reference-conditioned (Eq. 5): mean $\bar{c}$ moves $\leq 1$ pp under proposer swaps on GPQA ($0.716 \rightarrow 0.711$) and MedQA ($0.835 \rightarrow 0.827$), with transfer medians $0.767/0.767/0.768$ for $C$, vs. $0.696/0.663/0.659$ for $G$ and $0.627/0.620/0.617$ for $V$.

**Limitation: judge confound on $C$.** Because $C$ flows through an LLM judge, it could reflect judge linguistic priors. Two checks bound this: (i) on a 100-question MedQA spot-check, judge labels agree with exact-string matching on gold at Cohen's $\kappa{=}0.81$; string labels reduce $C$-only AUC by $0.02$–$0.04$ and preserve $C{+}G{+}V$ rankings. (ii) A judge-family swap (GPT-5-mini $\rightarrow$ Claude Sonnet 4.6) on two settings leaves $G$-only AUC unchanged within noise ($|\Delta| \leq 0.013$) and shifts cross-judge $C$-only AUC by $\leq 0.02$, preserving $C{+}G$ ordering (Appendix E.3). We note open-source judges remain untested.

**Multi-channel fusion.** $C{+}G{+}V$ beats the best single channel in $17/18$ combinations (median gain $+0.05$ AUC; bootstrap $95\%$ CI excludes zero in $12/18$; median fused AUC $0.780$, max $0.92$). Likelihood-ratio tests find incremental predictive value for $C$ conditional on $(G, V)$ in $18/18$ settings ($p \leq 0.002$), for $G$ conditional on $(C, V)$ in $16/18$, and for $V$ conditional on $(C, G)$ in $6/18$. Median AUROC is $0.670$ for $G$, $0.691$ for $G{+}V$, and $0.780$ for $C{+}G{+}V$, with full fusion improving over $G{+}V$ in $18/18$ settings (median $\Delta = +0.100$). Notably, $C{+}G$ achieves higher median AUROC ($0.793$) than $C{+}G{+}V$ (higher in $6/18$ settings), aligning with the LRT finding that $V$ contributes residual signal selectively. Residualizing on a leave-this-reasoner-out difficulty proxy leaves fused AUC $0.50$–$0.72$ (Appendix E.5). Verbalization correlates with accuracy across closed settings ($\rho{=}+0.71$, $N{=}6$) and contributes conditionally when internally stable, but adds little when the reasoner lacks metacognitive access. Thus, configuration choice should track deployment access: $G$ under black-box candidate access, $G{+}V$ with informative self-report, and $C{+}G{+}V$ when trusted verification is available.

Residualizing $C$, $G$, and $V$ on a leave-this-reasoner-out difficulty proxy still leaves fused AUC $0.50$–$0.72$ across the 18 combinations, so the three-channel signal is not reducible to shared question

[1] Stouffer's $z$ pools one-sided $p$-values via $z_{\text{pool}} = \sum_i z_i/\sqrt{n}$ with $z_i = \Phi^{-1}(1 - p_i)$.

difficulty. Appendix audits further show that the signal localizes to $\bar{c}$ and the top-2 cosine margin, and that a nonlinear second-stage model does not improve on the linear fusion (median $\Delta = -0.023$; Appendix E.5).

Verbalization remains a gated metacognitive channel rather than a universal pillar: $\mathrm{AUC}(V)$ correlates with reasoner accuracy across closed-set settings (Spearman $\rho = +0.71$, $N$=6), the LRT adds signal in only $6/18$ open-ended settings, and the MedQA / Gemini / Gemini-proposer counterexample shows that strong standalone $V$ need not lift $C$+$G$. The appropriate configuration therefore depends on available access: $G$ in candidate-defined black-box settings, $G$+$V$ when self-report is useful and passes setting-level calibration and orthogonality tests (Appendix E.5), and $C$+$G$+$V$ only when a trusted reference or verification signal is available (Section 5.3).

## 6 Conclusion

Geometry is this paper's primary novel, trace-derived confidence signal: it measures trajectory–candidate alignment from generated text alone, without requiring token probabilities, hidden states, or generator weights at inference. The reported headline result remains the *reference-assisted* fusion: $C$+$G$+$V$ at $K$=4 outperforms SC@8 in AUC across all six evaluated benchmark–reasoner settings, with median $\Delta\mathrm{AUC} = +0.075$ and pooled Stouffer $z = 5.67$ ($p = 7.1 \times 10^{-9}$); the gain survives a fixed-pick falsification ($+0.060$ median AUC at SC's own plurality answer, $94\%$ pick-agreement) and re-running with an external embedder (E5-large-v2: $6/6$, $z = 3.14$).

The channels have different access requirements. $G$ is candidate-conditional; $V$ is a reasoner self-report; and $C$ is an external, reference- or verifier-conditioned reachability signal. Controlled trajectory-swap and body/final-answer interventions support a pairing-specific, position-dependent *behavioural* interpretation of $G$, but do not identify the generator's hidden computational mechanism, and we make no causal claim about internal representations.

The practical implication is therefore not that one confidence estimator dominates universally. Against strong baselines on matched traces, Geometry is competitive with, rather than uniformly better than, well-elicited verbal confidence and aggregate token likelihood (Section 4.6). What the results support is that different access regimes admit different operating points: verbal confidence when simplicity dominates, $G$ or $G$+$V$ under black-box candidate-defined access, and full $C$+$G$+$V$ when trusted external verification is available.

An orthogonal route to calibrated black-box confidence is training models whose outputs are calibrated by construction rather than estimated post-hoc; concurrent industry work (System One Models / Jev, trained via reinforcement learning for calibrated decisions) explores this direction for structured decision tasks, trading free-form generation for typed, calibrated probabilistic outputs [Almeida, 2026].

**Limitations.** *(i) Coverage.* $C$ requires a reference answer or verifier and is not reasoner-intrinsic; it is unavailable in open-ended settings without verification. Reference-free reconstructions from pool morphology recover little signal. Reformulating reachability as a multiagent decision problem (e.g., cross-examination between proposer agents) is an open direction. *(ii) Candidate quality.* $G$ requires candidate anchors and cannot detect if the candidate set lacks viable answers (Appendix B); candidate-space adequacy is an external prerequisite. *(iii) Judge robustness.* Only one alternate judge across two settings was evaluated; open-source judges remain untested. *(iv) Non-causal interpretation.* Interventions are behavioral text manipulations, not causal claims about hidden representations; we withdraw the earlier terminal sign-flip account (Appendix C). *(v) Linguistic scope.* Benchmarks are English and embedders strong; multilingual transfer remains untested. *(vi) Agentic actions.* Extending candidate spaces to tool calls or actions is prospective and unevaluated. *(vii)* No within-trace score can recover an unproposed hypothesis.

## Acknowledgments and Disclosure of Funding

This work was funded by Novo Nordisk. The authors declare no competing interests.

## References


Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. Do language models know when they're hallucinating references? In *Findings of the association for computational linguistics: Eacl 2024*, pages 912–928, 2024.

Riccardo Ali, Francesco Caso, Christopher Irwin, and Pietro Liò. Entropy-lens: The information signature of transformer computations. *arXiv preprint arXiv:2502.16570*, 2025.

Diogo Almeida. Introducing system one models & jev. TypeSafe AI Blog, September 2026. URL `https://typesafe.ai/blog/introducing-system-one-models-and-jev`. Accessed 2026-09-30.

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*, pages 967–976, 2023.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Jeremy Cole, Michael Zhang, Dan Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Xinyan Jiang, Ninghao Liu, Di Wang, and Lijie Hu. Beyond scalars: Evaluating and understanding llm reasoning via geometric progress and stability. *arXiv preprint arXiv:2603.10384*, 2026.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

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

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, et al. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

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.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2303.08896.

Marc Boubnovski Martell, Kaspar Märtens, Lawrence Phillips, Daniel Keitley, Maria Dermit, and Julien Fauqueur. A scalable LLM framework for therapeutic biomarker discovery: Grounding Q/A generation in knowledge graphs and literature. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

Marc Boubnovski Martell, Josefa Lia Stoisser, Lawrence Phillips, Aditya Misra, Robert Kitchen, Jesper Ferkinghoff-Borg, Jialin Yu, Philip Torr, and Kaspar Märtens. Mechpert: Mechanistic consensus as an inductive bias for unseen perturbation prediction. *arXiv preprint arXiv:2602.13791*, 2026.

Tejaswini Pedapati, Amit Dhurandhar, Soumya Ghosh, Soham Dan, and Prasanna Sattigeri. Large language model confidence estimation via black-box access. *arXiv preprint arXiv:2406.04370*, 2024.

Edward Phillips, Sean Wu, Soheila Molaei, Danielle Belgrave, Anshul Thakur, and David Clifton. Geometric uncertainty for detecting and correcting hallucinations in llms. *arXiv preprint arXiv:2509.13813*, 2025.

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. *arXiv preprint arXiv:2311.12022*, 2023.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38, 2025.

Josefa Lia Stoisser, Marc Boubnovski Martell, Kaspar Märtens, Lawrence Phillips, Stephen Michael Town, Rory Donovan-Maiye, and Julien Fauqueur. Query, don't train: Privacy-preserving tabular prediction from ehr data via sql queries. *arXiv preprint arXiv:2505.21801*, 2025a.

Josefa Lia Stoisser, Marc Boubnovski Martell, Lawrence Phillips, Gianluca Mazzoni, Lea Mørch Harder, Philip Torr, Jesper Ferkinghoff-Borg, Kaspar Martens, and Julien Fauqueur. Towards agents that know when they don't know: Uncertainty as a control signal for structured reasoning. *arXiv preprint arXiv:2509.02401*, 2025b.

Josefa Lia Stoisser, Lawrence Phillips, Aditya Misra, Tom A Lamb, Philip Torr, Marc Boubnovski Martell, Julien Fauqueur, and Kaspar Märtens. Towards label-free biological reasoning synthetic dataset creation via uncertainty filtering. *arXiv preprint arXiv:2510.05871*, 2025c.

Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. LLM reasoning as trajectories: Step-specific representation geometry and correctness signals. *arXiv preprint arXiv:2604.05655*, 2026.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, 2023.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025.

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

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.14489.

# A Method and Pipeline Schematics

## A.1 Geometry Scoring over Reasoning Trajectories

Figure 1 gives a visual schematic of the Geometry ($G$) scoring pipeline described in Section 3: each sampled chain-of-thought is embedded as a sliding-window trajectory, the penultimate window is compared by cosine distance to each fixed candidate anchor, and the resulting per-option distances $D_j$ feed the one-parameter softmax $\textsc{Geo}(q)$.

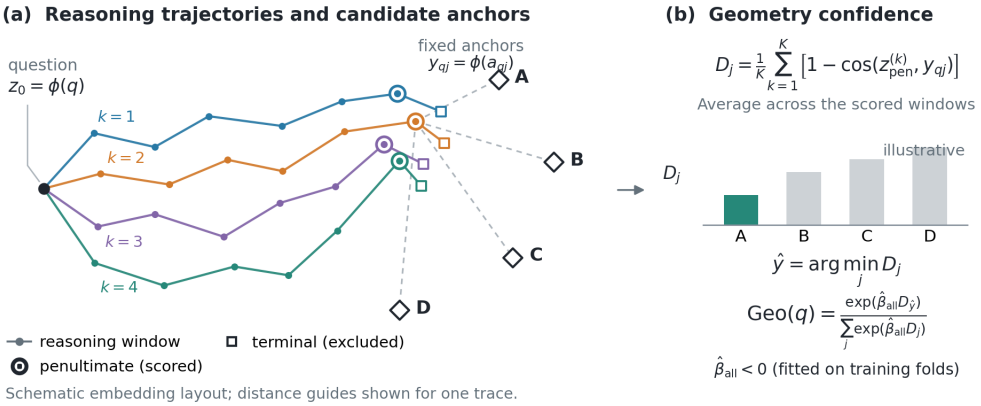



Figure 1: **Schematic illustration of the Geometry channel.** (a) Four sampled reasoning trajectories ($k$=1, . . . , 4) in embedding space, each a sliding-window path from the question embedding $z_0$ toward one of four fixed candidate anchors $A$–$D$ ($y_{qj} = \phi(a_{qj})$). Open squares mark the excluded terminal window; filled rings mark the scored penultimate window. (b) The penultimate-window distances $D_j$ (Eq. 1, mean cosine distance over the $K$ samples) are passed through the one-parameter softmax to obtain $\textsc{Geo}(q)$, with the predicted answer $\hat{y} = \arg\min_j D_j$. **Conceptual illustration, not measured data;** the embedding layout and distances shown are schematic, not an actual fitted projection.

## A.2 Three-Channel Pipeline: Proposer, Judge, and Reasoner

Figure 2 summarizes the agent roles and information flow of the open-ended three-channel protocol defined in Section 5.2: an independent proposer generates a hypothesis pool, a fixed judge establishes gold-equivalence to populate Coverage ($C$) features, the reasoner's own sampled trajectories are scored against proposer-hypothesis anchors to populate Geometry ($G$) features, and the reasoner's self-report on its own trace populates Verbalization ($V$) features. The three feature sets are then combined by the two-stage logistic stacking of Eq. 6.

# B Anchor-Design Ablation

The central design choice of our pipeline is what to embed as the per-option anchor against which we measure trajectory convergence. Table 5 compares three candidates on GPQA Diamond under an identical open-ended protocol (CoT generated without options shown) and identical trajectory construction (3072-dim Gemini embeddings, $W$=30, $S$=15):

**Condensed-answer anchor.** For each of the $K$=4 samples, extract the model's final answer phrase and use its embedding as the anchor for that sample's trajectory. Because models often produce the same short answer string across samples, the four per-sample anchors collapse to the same point and the geometric "choice" becomes degenerate, reflected in near-chance Geo AUC (0.44 Gemini, 0.32 Sonnet).

**Reasoning-tail anchor.** Use the embedding of the last ∼2000 characters of each sample's own CoT as the anchor. This avoids the collapse of the condensed-answer variant but introduces a different pathology: the anchor lives inside the trajectory it is scoring, so final-window distance is trivially small for every option. Geo AUC stays near chance (0.48/0.45).

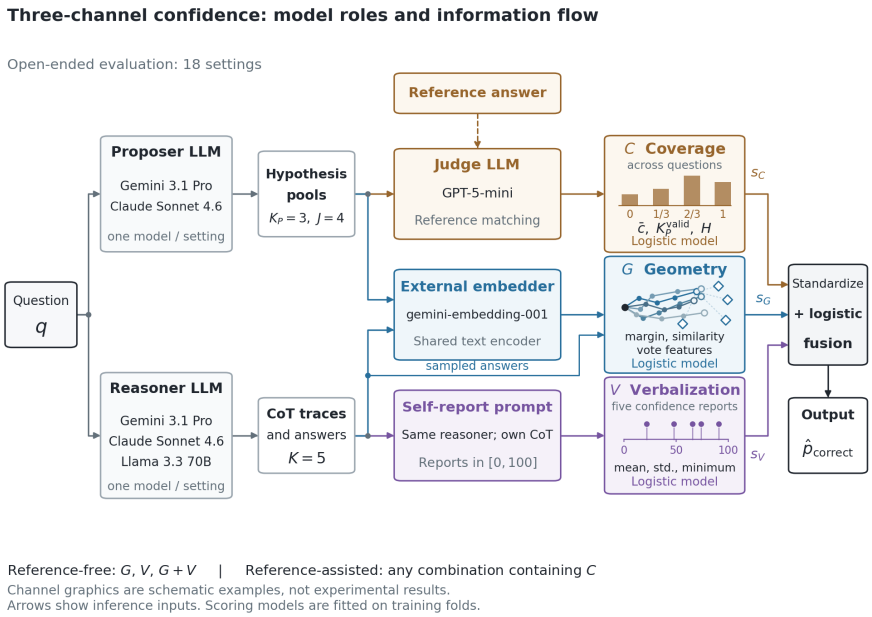



Figure 2: **Three-channel pipeline overview.** Model roles (reasoner, proposer, judge) and the resulting Coverage ($C$), Geometry ($G$), and Verbalization ($V$) feature streams, as defined in Section 5.2 and fused via Eq. 6. $C$ is reference-conditioned (requires the judge's gold-equivalence label); $G$ is candidate-conditional but verifier-free; $V$ is reasoner-intrinsic (Section 5.3).

Table 5: **Anchor-design ablation on GPQA Diamond.** All three variants use the same trajectory and the same softmax scoring; only the per-option anchor differs. Geo AUC measures selective prediction from the single-feature softmax probability; PickAcc is top-1 accuracy of the argmin-distance rule.

| **Anchor design** | **Model** | $N$ | Geo AUC | Composite | $\Delta$(Geo) |
|---|---|---|---|---|---|
| Answer (condensed) | Gemini 3.1 Pro | 117 | 0.437 | 0.520 | n/a |
| Reasoning tail | Gemini 3.1 Pro | 117 | 0.479 | 0.545 | +0.006 |
| **Dataset option (ours)** | Gemini 3.1 Pro | 117 | **0.710** | **0.701** | +0.141 |
| Answer (condensed) | Claude Sonnet 4.6 | 187 | 0.319 | 0.664 | n/a |
| Reasoning tail | Claude Sonnet 4.6 | 187 | 0.450 | 0.473 | +0.011 |
| **Dataset option (ours)** | Claude Sonnet 4.6 | 187 | **0.750** | **0.727** | +0.122 |

**Dataset option anchor (ours).** Use the embedding of each of the four official benchmark option texts as fixed, external anchors. These are identical across all samples, independent of model output, and carry domain-relevant semantic content. Geo AUC rises to $0.710/0.750$.

The ablation suggests a structural lesson for this setup: anchors derived from the model's own output are either close to a point estimate of its answer or closely tied to the reasoning trace being scored. By contrast, an external dataset-supplied anchor provides the cleanest reference point against which trajectory convergence can be measured in our experiments.

This is also the empirical basis for treating candidate quality as a first-class deployment limitation (Section 5.3). Geometry measures commitment relative to the candidate set it is given; it cannot determine by itself whether that set contains a viable answer. The gap between near-chance model-derived anchors (0.32–0.48) and dataset options (0.710–0.750) is large enough that candidate-space adequacy dominates the design, and it is an external requirement rather than something $G$ solves. We note for completeness that we do not report a gold-absent stress test here: removing the gold option from a four-option set leaves three options, so the softmax normalisation and the fitted $\beta$ are no longer comparable across conditions, and the resulting confound would make such a test uninterpretable without an equal-cardinality distractor substitution.

Table 6: **Calibration in the main blinded option-anchor protocol.** A strongly negative $\hat{\beta}$ means closer option anchors are more likely to be correct. Disc. AUC is single-feature selective-prediction AUC of GEO.

| Dataset | Model | $\hat{\beta}$ | $z$ | Disc. AUC | $N$ |
|---|---|---|---|---|---|
| MedQA | Gemini 3.1 Pro | $-25.064$ | $-25.39^{***}$ | 0.821 | 1272 |
| MedQA | Claude Sonnet 4.6 | $-22.109$ | $-23.81^{***}$ | 0.753 | 1271 |
| GPQA | Gemini 3.1 Pro | $-25.315$ | $-5.48^{***}$ | 0.710 | 117 |
| GPQA | Claude Sonnet 4.6 | $-30.491$ | $-7.44^{***}$ | 0.750 | 187 |

$^{***}p < 10^{-6}$, $^{**}p < 10^{-3}$, $^{*}p < 0.05$

Table 7: **Separate robustness evaluation with committed-rationale anchors and a third benchmark.** $\tilde{D}$ is the null-standardized top-two margin from independently generated committed-rationale anchors; Composite adds an option-text tail-margin feature from the original trajectories. This evaluation uses a smaller subset, $K{=}5$, and the target "the plurality-vote pick is correct," so the AUCs are not directly comparable to Table 1. $\rho$ is the Pearson correlation between $\tilde{D}$ and the option-text tail margin.

| Benchmark | $N$ | $\tilde{D}$ alone | Composite | $\Delta$ vs $\tilde{D}$ | $\rho(\tilde{D}, \text{tail})$ |
|---|---|---|---|---|---|
| MedQA-USMLE | 197 | 0.862 | **0.889** | $+0.027$ | $+0.38$ |
| GPQA Diamond | 194 | 0.753 | **0.782** | $+0.029$ | $+0.39$ |
| MMLU-Pro | 191 | 0.722 | **0.779** | $+0.057$ | $+0.37$ |

### B.1 Main blinded option-anchor calibration

### B.2 Committed-rationale anchors and a third benchmark

A distinct alternative is to use, for each option, an independently generated *committed rationale*: a short paragraph asserting that option as correct, produced by a separate anchor model under the same blinded-generation regime. On MedQA, GPQA, and MMLU-Pro we summarize the resulting per-question option scores by a null-standardized top-two margin $\tilde{D}$ and ask whether it adds rank information beyond an option-text tail margin computed from the original trajectories. Because this evaluation uses a smaller subset, $K{=}5$, and the binary target "the plurality-vote pick is correct," we report it here rather than in the main comparison table. The qualitative result is stable: the committed-rationale score is discriminative on all three benchmarks, and combining it with the option-text tail margin improves AUC by $+0.027$ on MedQA, $+0.029$ on GPQA, and $+0.057$ on MMLU-Pro.

**$\hat{\beta}$ on a third benchmark.** As a direct robustness check on the Section 4.2 finding that closer anchors are more likely correct, we refit the one-parameter softmax on this same committed-rationale-anchor pipeline across all three benchmarks. The covariate is the per-option mean cosine *distance* from the $K{=}5$ trajectory tails to the committed-rationale anchors, and the choice is the gold option. Anchor counts are heterogeneous on MMLU-Pro ($J \in \{4, \ldots, 10\}$ per question, mostly $J{=}10$), which the softmax formulation naturally accommodates by computing each question's softmax over its own option set. The $\hat{\beta}$ values (Table 8) are strongly negative and highly significant on every benchmark, including MMLU-Pro at $N{=}200$, providing third-benchmark replication of the calibration result. The numerical magnitudes are not directly comparable to Table 6 because the anchors differ (committed rationales vs. raw option text); the comparable claim is the *sign and significance* of $\hat{\beta}$.

### B.3 Budget-matched self-consistency on GPQA and MMLU-Pro

For symmetry, we also report budget-matched ($K{=}4$) self-consistency on GPQA Diamond and on the MMLU-Pro pilot, using the same character-bigram Jaccard mapping rule (GPQA) or the embedding cluster mode (MMLU-Pro) already used by the pipeline. The geometric signal exceeds $K{=}4$ self-consistency in five of six MMLU-Pro combinations and on GPQA/Gemini; GPQA/Sonnet is the one near-tie setting, where self-consistency reaches $0.747$ versus our $0.727$ AUC.

Table 8: $\hat{\beta}$ **under the committed-rationale-anchor pipeline.** Same softmax objective as Table 6 but with committed-rationale (Sonnet 4.6) anchors and Gemini 3.1 Pro $K{=}5$ blinded CoT trajectories. $\hat{\beta}$ is strongly negative on every benchmark, including MMLU-Pro with heterogeneous $J \in \{4, \ldots, 10\}$ per question.

| **Benchmark** | $N$ | $\hat{\beta}$ | SE | $z$ |
|---|---|---|---|---|
| GPQA Diamond | 198 | $-70.60$ | 7.45 | $-9.48^{***}$ |
| MedQA-USMLE | 200 | $-64.38$ | 6.63 | $-9.71^{***}$ |
| MMLU-Pro | 200 | $-51.77$ | 4.20 | $-12.34^{***}$ |

$^{***}p < 10^{-6}$

Table 9: **Self-consistency comparison in the cleanest composite settings.** MedQA rows use the matched $N{=}273$ held-out split from the self-consistency audit and compare against the stronger SC@10 baseline; GPQA rows use budget-matched SC@4 on the retained subsets of Table 1. Positive $\Delta$AUC means the geometric composite outperforms self-consistency.

| **Benchmark** | **Model** | **SC baseline** | **Our AUC** | $\Delta$AUC |
|---|---|---|---|---|
| MedQA (held-out) | Gemini 3.1 Pro | SC@10 = 0.643 | **0.827** | +0.184 |
| MedQA (held-out) | Claude Sonnet 4.6 | SC@10 = 0.712 | **0.759** | +0.047 |
| GPQA | Gemini 3.1 Pro | SC@4 = 0.690 | **0.701** | +0.011 |
| GPQA | Claude Sonnet 4.6 | **SC@4 =** 0.747 | 0.727 | −0.020 |

## C Window-Position Sweep and the Endpoint-Definition Correction

To test whether the softmax signal is a property of the question stem (which would yield a flat $\hat{\beta}$ across window positions) or of the trajectory's late-stage convergence, we refit the one-parameter softmax of Section 3 after replacing each trace's penultimate-window aggregate with a single sliding window at position $k$ from the end of the trace. We sweep $k \in \{-1, -2, \ldots, -8\}$ on GPQA Diamond ($N{=}197$ questions, $K{=}5$ samples per question, $W{=}30$, $S{=}15$) for both Gemini 3.1 Pro and Claude Sonnet 4.6, with 95% bootstrap confidence intervals over 200 resamples. We additionally replicate the sweep on MedQA-USMLE ($N{=}200$, $K{=}5$) and MMLU-Pro ($N{=}200$, $K{=}5$) using newly built sliding-window caches under the identical protocol; results are reported in Table 12.

**Endpoint-definition artifact and its correction.** An earlier version of this appendix reported a *positive* coefficient at the terminal position $k{=}-1$ on GPQA ($+22.1$ Gemini, $+23.7$ Claude) and interpreted it as evidence of answer commitment. That value did not come from a natural final CoT window. It was computed from an *appended bare answer label*: a short synthetic string of the form "The answer is (B)" concatenated to the trace. A natural final CoT window and an appended symbolic answer label are different objects, and only the former is produced by the reasoner in the deployment setting we study. We document the correction here because the distinction is easy to lose in a sliding-window pipeline.

Table 11 reports the corrected values together with three controls. On natural CoT windows the terminal coefficient is strongly *negative* on both generators, and the penultimate coefficient is more negative still, so the previously reported terminal reversal disappears. Replacing the bare label with the corresponding *full option text* also gives strongly negative coefficients, and using the gold option text as the endpoint gives the extreme value expected of a near-oracle endpoint. Decisively, appending a *constant* label string, identical for every question and therefore carrying no question-specific answer information, reproduces a positive coefficient of similar magnitude ($\approx +20$ to $+29$). A quantity that a constant string can reproduce cannot be evidence that the trajectory encodes the model's committed answer. We therefore withdraw the answer-commitment interpretation of the terminal position, along with the associated prediction that hidden-state probes should show a sign flip at the same position.

**What the corrected sweep does and does not show.** Two conclusions survive. First, $\hat{\beta}$ is *not* flat across $k$, which rules out the interpretation that the calibration signal is a question-stem effect that any tail window would inherit. Second, the coefficient is deepest-negative in the late-but-not-terminal region, and the penultimate-window trough holds in $4/4$ MedQA / MMLU-Pro combinations

Table 10: **Budget-matched self-consistency ($K$=4) vs. our geometric score on GPQA and MMLU-Pro.** Self-consistency confidence = majority fraction over $K$=4 samples, mapped to options by character-bigram Jaccard (GPQA) or embedding cluster mode (MMLU-Pro). "Our AUC" = composite (Table 1) for GPQA, and gated geometric AUC (per-setting 5-fold CV) for MMLU-Pro. MMLU-Pro $N$ matches Section 3.1 ($N$=200 per setting, $N$=199 for prop / Gemini).

| Benchmark | Cond. | Model | SC $K$=4 Acc | SC $K$=4 AUC | Our AUC |
|---|---|---|---|---|---|
| GPQA Diamond | — | Gemini 3.1 Pro | 0.778 | 0.690 | **0.701** |
| GPQA Diamond | — | Claude Sonnet 4.6 | 0.699 | **0.747** | 0.727 |
| MMLU-Pro | MCQ | Gemini 3.1 Pro | 0.915 | 0.648 | **0.690** |
| MMLU-Pro | MCQ | Claude Sonnet 4.6 | 0.890 | 0.590 | **0.623** |
| MMLU-Pro | open | Gemini 3.1 Pro | 0.775 | 0.540 | **0.562** |
| MMLU-Pro | open | Claude Sonnet 4.6 | 0.745 | 0.575 | **0.623** |
| MMLU-Pro | prop | Gemini 3.1 Pro | 0.648 | 0.629 | **0.650** |
| MMLU-Pro | prop | Claude Sonnet 4.6 | 0.650 | 0.567 | **0.639** |

Table 11: **Endpoint-definition correction on GPQA Diamond.** $\hat{\beta}_{\text{all}}$ under different definitions of the trajectory endpoint, with 95% bootstrap CIs where available. Natural CoT windows give negative coefficients at both terminal and penultimate positions. The positive value previously reported is reproduced by an appended *constant* label string that contains no question-specific information, identifying it as a property of the appended label rather than of answer commitment.

| Endpoint definition | Reasoner | $\hat{\beta}_{\text{all}}$ | 95% CI |
|---|---|---|---|
| Natural CoT window, $k$=−2 (ours) | Gemini 3.1 Pro | −48.2 | [−66.4, −34.2] |
| Natural CoT window, $k$=−2 (ours) | Claude Sonnet 4.6 | −39.5 | [−54.2, −27.8] |
| Natural CoT window, $k$=−1 | Gemini 3.1 Pro | −42.6 | [−54.5, −33.0] |
| Natural CoT window, $k$=−1 | Claude Sonnet 4.6 | −28.8 | [−41.6, −18.4] |
| Predicted-option full text | Gemini 3.1 Pro | −47.4 | — |
| Predicted-option full text | Claude Sonnet 4.6 | −36.5 | — |
| Gold-option full text (oracle) | Gemini 3.1 Pro | −111.2 | — |
| Gold-option full text (oracle) | Claude Sonnet 4.6 | −111.2 | — |
| Appended bare answer label *(previously reported)* | Gemini 3.1 Pro | +21.9 | [+7.2, +38.2] |
| Appended bare answer label *(previously reported)* | Claude Sonnet 4.6 | +23.5 | [+8.8, +40.3] |
| Appended *constant* label (control) | Gemini 3.1 Pro | +20.7 to +28.9 | — |
| Appended *constant* label (control) | Claude Sonnet 4.6 | +20.1 to +29.6 | — |

(Table 12); on GPQA the corrected natural coefficients are also more negative at $k$=−2 than at $k$=−1 on both generators, consistent with Figure 3. We therefore retain the penultimate natural window as the scoring position, on the design rationale that it avoids explicit answer-endpoint content, and we no longer claim a positive natural terminal coefficient or a benchmark-specific terminal reversal. The forced-label experiments are retained above only as controls that identify the artifact.

Table 12: **Window-position replication on MedQA-USMLE and MMLU-Pro.** Aggregate $\hat{\beta}_{\text{all}}$ at $k$=−1 and $k$=−2 on natural CoT windows. Both positions are strongly negative on both benchmarks, and the penultimate position is at or near the trough.

| Benchmark | Reasoner | $N$ | acc($k$=−1) | $\hat{\beta}_{\text{all}}(k$=−1) | $\hat{\beta}_{\text{all}}(k$=−2) |
|---|---|---|---|---|---|
| MedQA-USMLE | Gemini 3.1 Pro | 200 | 0.88 | −65.4 | −67.1 |
| MedQA-USMLE | Claude Sonnet 4.6 | 199 | 0.80 | −62.7 | −57.8 |
| MMLU-Pro | Gemini 3.1 Pro | 200 | 0.55 | −52.6 | −56.5 |
| MMLU-Pro | Claude Sonnet 4.6 | 200 | 0.52 | −49.1 | −50.0 |

**Per-cell body/final-answer intervention.** Table 13 gives the per-cell values behind the body/final-answer intervention of Section 4.4, with 95% percentile bootstrap CIs over 2,000 question resamples. The pattern is consistent in all four cells: replacing the body collapses pick accuracy, restoring the

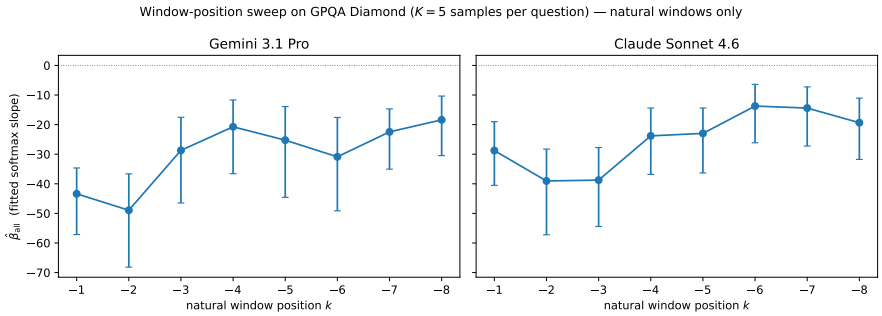



Figure 3: **Corrected window-position sweep on GPQA Diamond.** $\hat{\beta}_{\text{all}}$ as a function of natural CoT window position $k$ (no appended answer label at any position), for Gemini 3.1 Pro and Claude Sonnet 4.6, with $95\%$ bootstrap CIs over 200 resamples. All eight positions are negative for both generators; the deepest trough is in the late-but-not-terminal region ($k{=}-2/k{=}-3$), matching the natural-window rows of Table 11. There is no terminal-position sign reversal once the endpoint is a genuine trailing CoT window rather than an appended "The answer is (X)" label; the positive coefficient reported in an earlier version of this appendix is reproduced only by that label artifact (see correction above and Table 11).

original final answer sentence recovers it, and an appended counterfactual wrong answer captures the terminal window.

Table 13: **Body/final-answer intervention, per cell.** Candidate-pick accuracy for each constructed window, with $95\%$ bootstrap CIs. "CF wrong" reports the rate at which the appended counterfactual *wrong* answer is selected (higher means terminal geometry follows the appended answer string). Body swaps use the nearest-neighbour body from another question, the hardest available substitution.

| **Cell** | $N$ | Body (orig.) | Body (swapped) | Swap+orig. final | Swap+swap. final | CF wrong |
|---|---|---|---|---|---|---|
| GPQA / Gemini | 197 | 0.528 [0.457, 0.599] | 0.305 [0.244, 0.371] | 0.553 [0.482, 0.624] | 0.320 [0.259, 0.386] | 0.777 [0.721, 0.838] |
| GPQA / Claude | 197 | 0.477 [0.406, 0.548] | 0.340 [0.274, 0.406] | 0.442 [0.376, 0.508] | 0.284 [0.223, 0.350] | 0.777 [0.716, 0.832] |
| MedQA / Gemini | 200 | 0.720 [0.655, 0.785] | 0.450 [0.385, 0.520] | 0.755 [0.695, 0.815] | 0.405 [0.340, 0.475] | 0.895 [0.850, 0.935] |
| MedQA / Claude | 196 | 0.719 [0.653, 0.781] | 0.403 [0.337, 0.474] | 0.709 [0.648, 0.770] | 0.347 [0.281, 0.413] | 0.832 [0.781, 0.883] |
| **Median** | — | **0.624** | 0.372 | 0.631 | 0.333 | 0.804 |

## D Third-Generator Scaling: Open-Weight Llama 3.3 70B

To test whether the trajectory-geometry phenomenon is an artifact of the two frontier closed-API generators in the main paper (Gemini 3.1 Pro and Claude Sonnet 4.6) or a property of CoT reasoning that scales with capability, we replicate the $\hat{\beta}$ and GEO AUC measurements with an open-weight third-family generator: **Meta Llama 3.3 70B Instruct**. The protocol is identical to the committed-rationale robustness pipeline above: $K{=}5$ blinded rationales at $T{=}0.7$, embedded with `gemini_embedding`, and scored against the same Sonnet 4.6 committed-rationale anchors used in Tables 7 and 8. Results are reported in Table 14.

Two readings of the table are worth making explicit. (a) The *direction* of the effect is universal across the model spectrum: every setting of every (benchmark, generator) combination in the paper, spanning a $2\times$ accuracy range and three model families (Google Gemini 3.1 Pro, Anthropic Claude Sonnet 4.6, Meta Llama 3.3 70B), produces a strongly negative $\hat{\beta}$. The trajectory-drift phenomenon is not a frontier-only or closed-API-only effect. (b) The *magnitude* of the usable selective-prediction signal scales with model capability: on MedQA where Llama remains a competent solver (acc $0.66$), GEO AUC is $0.76$, comparable to the published frontier combinations; on GPQA where Llama is at near-random ($0.33$), the geometric channel naturally has little to discriminate, and GEO AUC drops to $0.54$. This is the expected behavior under our channel-limit interpretation: the geometry can amplify a model's existing convergence into a calibrated confidence score, but cannot manufacture confidence where the model has none.

Table 14: Third-generator scaling: open-weight Llama 3.3 70B Instruct paired with the same Sonnet 4.6 option anchors used for the published Gemini combinations. Top-1 accuracy is the rate at which the geometric $\arg\min$ matches the gold option (closed-set discrimination from open-ended generation). Across all three benchmarks $\hat{\beta}$ remains strongly negative ($z \leq -2.4$), confirming that answer-directed trajectory drift is not a frontier-API artifact. GEO AUC tracks model capability: it remains useful on MedQA and MMLU-Pro, where Llama is competitive; on GPQA, where Llama's accuracy collapses to near-chance (0.33), GEO AUC also degrades to near-chance, as the geometric channel cannot extract a confident-correct signal from a model that has not converged to the correct option.

| combination | $N$ | top-1 acc | $\hat{\beta}$ [95% CI] | $z$ | GEO AUC [95% CI] |
|---|---|---|---|---|---|
| GPQA / Llama 3.3 70B | 198 | 0.328 | $-6.95$ $[-11.87, -1.42]$ | $-2.4$ | 0.541 [0.458, 0.620] |
| MedQA / Llama 3.3 70B | 200 | 0.665 | $-23.70$ $[-29.09, -19.47]$ | $-9.6$ | 0.757 [0.686, 0.819] |
| MMLU-Pro / Llama 3.3 70B | 200 | 0.305 | $-11.37$ $[-16.14, -6.91]$ | $-4.7$ | 0.626 [0.542, 0.711] |

Table 15: Full proposer-level three-channel selective-prediction AUC in open-ended MCQA across three benchmarks $\times$ three reasoners $\times$ two proposers ($N \approx 200$ per cell). **Bold** marks the best single-channel value among $C/G/V$; *italic* marks rows where $C{+}G{+}V$ exceeds every two-channel ablation.

| Bench | Reasoner | Proposer | $N$ | C | G | V | C+G | G+V | C+V | C+G+V |
|---|---|---|---|---|---|---|---|---|---|---|
| GPQA | Gemini | Claude | 193 | **0.81** | 0.66 | 0.67 | 0.83 | 0.71 | 0.80 | *0.82* |
| GPQA | Gemini | Gemini | 195 | **0.83** | 0.81 | 0.71 | 0.92 | 0.81 | 0.83 | *0.92* |
| GPQA | Claude | Claude | 193 | **0.76** | 0.68 | 0.67 | 0.79 | 0.70 | 0.78 | *0.81* |
| GPQA | Claude | Gemini | 194 | 0.74 | **0.80** | 0.75 | 0.88 | 0.83 | 0.82 | *0.89* |
| GPQA | Llama | Claude | 193 | **0.64** | 0.59 | 0.62 | 0.62 | 0.64 | 0.68 | *0.68* |
| GPQA | Llama | Gemini | 195 | **0.72** | 0.68 | 0.49 | 0.82 | 0.66 | 0.69 | *0.82* |
| MedQA | Gemini | Claude | 199 | **0.75** | 0.72 | 0.58 | 0.85 | 0.73 | 0.80 | *0.86* |
| MedQA | Gemini | Gemini | 200 | **0.74** | 0.63 | 0.59 | 0.80 | 0.67 | 0.79 | 0.77 |
| MedQA | Claude | Claude | 199 | 0.72 | 0.62 | **0.78** | 0.81 | 0.77 | 0.86 | *0.86* |
| MedQA | Claude | Gemini | 199 | **0.74** | 0.69 | 0.74 | 0.84 | 0.76 | 0.84 | *0.86* |
| MedQA | Llama | Claude | 199 | **0.74** | 0.64 | 0.57 | 0.80 | 0.63 | 0.74 | 0.79 |
| MedQA | Llama | Gemini | 200 | **0.71** | 0.64 | 0.58 | 0.75 | 0.64 | 0.73 | *0.76* |
| MMLU-Pro | Gemini | Claude | 200 | 0.70 | **0.71** | 0.58 | 0.75 | 0.69 | 0.71 | 0.73 |
| MMLU-Pro | Gemini | Gemini | 200 | **0.74** | 0.69 | 0.56 | 0.79 | 0.69 | 0.73 | 0.77 |
| MMLU-Pro | Claude | Claude | 200 | 0.59 | **0.63** | 0.62 | 0.64 | 0.65 | 0.63 | *0.65* |
| MMLU-Pro | Claude | Gemini | 200 | 0.68 | **0.70** | 0.67 | 0.75 | 0.70 | 0.73 | 0.75 |
| MMLU-Pro | Llama | Claude | 200 | **0.70** | 0.58 | 0.51 | 0.69 | 0.51 | 0.68 | 0.67 |
| MMLU-Pro | Llama | Gemini | 200 | 0.64 | **0.64** | 0.54 | 0.70 | 0.62 | 0.64 | 0.67 |
| *Median across 18 cells* | | | | 0.72 | 0.66 | 0.61 | 0.80 | 0.69 | 0.74 | **0.78** |

# E Robustness Audits

## E.1 Full proposer-level three-channel panel

## E.2 Cross-embedder audit

A natural concern is that the geometric channel is partly induced by the particular external embedder. We therefore replace the 3072-dimensional Gemini embedding model with an independently trained E5-large-v2 encoder and rerun the matched $K{=}4$ vs. SC@8 protocol on the same $N{\approx}200$-per-combination subset on which the K=8 SC traces are available ($N \in \{196, 197, 199, 200, 200, 200\}$ across the six closed-set settings), without changing any downstream fitting, cross-validation, or bootstrap procedure. The Geo-only comparison softens but survives in aggregate: Geo@4 exceeds SC@8 with median $\Delta\text{AUC} = +0.027$, mean $+0.022$, and pooled one-sided Stouffer $z{=}1.73$ ($p{=}0.042$). More importantly, the central three-channel claim is unchanged under the swap: the E5-based $G{+}C{+}V$ fusion is positive in all six closed-set settings, with median $\Delta\text{AUC} = +0.035$ over SC@8 and pooled Stouffer $z{=}3.14$ ($p{=}8.5 \times 10^{-4}$). This is the right level of invariance to demand. The

absolute strength of the geometric channel depends on representation quality, but the qualitative ranking and the fusion advantage do not collapse when the embedder family changes.

Table 16: External-embedder audit with E5-large-v2. Positive entries mean the alternative score beats SC@8 on the same setting. Swapping the encoder weakens some Geo-only settings but leaves the pooled Geo effect positive and the fusion effect positive in all six combinations.

| combination | $\Delta$(GEO@4 − SC@8) | $\Delta$(G+C+V − SC@8) |
|---|---|---|
| MedQA / Gemini | $+0.037$ | $+0.057$ |
| MedQA / Claude | $+0.026$ | $+0.073$ |
| GPQA / Gemini | $-0.018$ | $+0.014$ |
| GPQA / Claude | $+0.029$ | $+0.027$ |
| MMLU-Pro / Gemini | $+0.017$ | $+0.026$ |
| MMLU-Pro / Claude | $+0.039$ | $+0.043$ |

### E.3 Judge-family swap

We directly test whether the Coverage channel inherits its predictive value from one judge family's semantic priors. Reusing the cached (gold, hypotheses) tuples for two combinations (MedQA / Gemini-proposer, $N$=200; GPQA / Gemini-proposer, $N$=192 aligned), we re-judge every (question, seed) tuple with Anthropic Claude Sonnet 4.6 in place of the original Azure-hosted `openai_gpt5_mini`, holding the prompt, parser, and downstream pipeline fixed. We then recompute $\bar{c}$ from the swapped labels and re-evaluate three quantities under both judges: $C$-only AUC, $C$+$G$ fused AUC (5-fold CV logistic on standardized features), and $G$-only AUC (judge-invariant feature, judge-dependent outcome). To separate *same-judge consistency* from genuine cross-judge stability we additionally report a cross-judge AUC: the swapped-judge $\bar{c}_{\text{swap}}$ scored against the *original* judge's outcome label, which is the comparison a skeptical reviewer should care about.

Table 17 summarises the result. Inter-judge agreement is high (Cohen's $\kappa$=0.82 on the 5-class MedQA labels, 0.72 on the binary coverage projection, raw agreement 90.5%), with the bulk of the disagreement being Claude rejecting coverage that GPT-mini accepted (15 of 19 MedQA disagreements), as expected for a stricter judge. Three patterns hold across both settings. (a) $G$-only AUC is unchanged within noise ($\Delta$=$-0.013$ on MedQA and $-0.001$ on GPQA), consistent with the within-trace geometric feature being insensitive to judge identity in these two settings. (b) Cross-judge $C$-only AUC changes by at most 0.02 in both settings (MedQA $+0.019$, GPQA $-0.017$), so the rank information in $\bar{c}$ does not collapse when the judge changes. (c) The $C$+$G$ ordering is preserved in both settings (no flip; same-judge $\Delta$= + 0.100, +0.006). In summary: *in the two tested judge-swap settings, cross-judge $C$-only AUROC changes by at most* 0.02 *while the $C$+$G$ ordering is preserved.* This is bounded evidence of judge-family robustness, not a demonstration of judge-independence; substantially weaker, open-source, or structurally different judges and the stability of the fusion weights remain untested. The same-judge $C$-only $\Delta$= + 0.131 on MedQA is partly a base-rate artifact (accuracy shifts 0.65→0.55 under the stricter judge), and we report the cross-judge number as the headline measurement; the same-judge column is included as a consistency check.

We stop short of claiming judge-independence: this is two combinations, one swap judge, and a stricter alternate. A full audit across all 18 combinations, additional judge families, weaker and open-source judges, fusion-weight stability, and human adjudication is left as future work. The check we run does, however, bound the strong form of the critique that $C$ is predominantly a GPT-family-prior signal.

### E.4 System cost of three-channel confidence

The strongest practical baseline in our setting is not SC@8 but higher-budget self-consistency. We therefore cost the three-channel system at $K$=4 against SC@10 using public May-2026 API prices for the two closed-source reasoners and the embedding model.[2] Table 18 reports the six-setting accounting. Pooled over the panel, $C$+$G$+$V$ at $K$=4 reduces total spend from \$0.2839 to \$0.1379

[2]Prices used (USD per 1M tokens, input / output): Gemini 3.1 Pro \$2 / \$12; Claude Sonnet 4.6 \$3.3 / \$16.5; `openai_gpt5_mini` (judge) \$0.25 / \$2; `gemini_embedding` \$0.15 / n/a. Sourced from the corresponding Vertex, AWS, and Azure model catalog pages, May 2026; archived snapshots are retained by the authors.

Table 17: Judge-family swap (Azure `openai_gpt5_mini` → Anthropic Claude Sonnet 4.6) on two open-ended settings. Cross-judge $\Delta$ scores the swapped-judge feature against the original-judge outcome label, isolating rank stability from judge-coupled base-rate motion. $G$-only $\Delta$ uses a judge-invariant feature with the judge-dependent outcome.

| Metric | MedQA / Gemini-prop | GPQA / Gemini-prop |
|---|---|---|
| $N$ aligned | 200 | 192 |
| Cohen's $\kappa$ (5-class) | 0.82 | — |
| Cohen's $\kappa$ (binary coverage) | 0.72 | — |
| Raw agreement (5-class) | 0.91 | — |
| $C$-only AUC, orig judge | 0.712 | 0.807 |
| $C$-only AUC, swap judge (same-judge) | 0.843 | 0.842 |
| $C$-only AUC, cross-judge ($c_{\text{swap}}$ vs. $y_{\text{orig}}$) | 0.731 | 0.790 |
| **Cross-judge $\Delta$ vs. orig** | $\mathbf{+0.019}$ | $\mathbf{-0.017}$ |
| $G$-only AUC ($y_{\text{orig}}$) | 0.637 | 0.764 |
| $G$-only AUC ($y_{\text{swap}}$) | 0.624 | 0.763 |
| $G$-only $\Delta$ | $-0.013$ | $-0.001$ |
| $C{+}G$ fused AUC, orig judge | 0.795 | 0.916 |
| $C{+}G$ fused AUC, swap judge | 0.895 | 0.922 |

per question, a $51.4\%$ saving, with per-setting savings of $42.6$–$54.4\%$. CoT generation remains the dominant line item inside the composite system (70–88% of total cost depending on setting); proposer, judge, and verbalization calls are second-order; embeddings are negligible (about $\$3 \times 10^{-5}$ per question). We emphasise that this monetary comparison is against SC@10, whereas the AUC comparison in Section 5.4 is against SC@8; the two should be read as separate statements, and we do not claim a matched-cost dominance over SC@8.

Table 18: Per-question system cost under public May-2026 API prices. SC@10 is ten-sample self-consistency; $G{+}C{+}V@4$ is the three-channel system with four reasoner samples plus proposer, judge, verbalization, and embedding calls.

| Combination | SC@10 | $G{+}C{+}V@4$ | Saving |
|---|---|---|---|
| MedQA / Gemini | \$0.0202 | \$0.0111 | 45.0% |
| MedQA / Claude | \$0.0607 | \$0.0294 | 51.6% |
| GPQA / Gemini | \$0.0330 | \$0.0162 | 50.9% |
| GPQA / Claude | \$0.0914 | \$0.0416 | 54.4% |
| MMLU-Pro / Gemini | \$0.0173 | \$0.0099 | 42.6% |
| MMLU-Pro / Claude | \$0.0613 | \$0.0296 | 51.7% |

**Performance–access–complexity frontier.** Monetary cost is only one axis. Table 19 summarises the operating points of Section 5.3 along the axes a practitioner actually has to satisfy: how many calls of each type are needed, whether token-level probabilities are required, whether a per-question reference or trusted verifier is required, and what performance is obtained. The table makes the central trade-off explicit. Verbal-only confidence is by far the simplest configuration and is already competitive on the aligned GPT-4.1 comparison. $G$ and $G{+}V$ add an embedding pass and a candidate-anchor requirement but no reference. Full $C{+}G{+}V$ is the strongest configuration we report, but it is the only one that requires a trusted reference or verifier, and it adds proposer and judge calls. Privileged likelihood baselines require token-probability access that several frontier endpoints do not expose.

### E.5 Verbalization sensitivity

The asymmetric LRT result for $V$ suggests that verbalization is neither uniformly redundant nor uniformly useful. We therefore measure, on the six closed-set settings, how the standalone verbalized-confidence AUC and the incremental gain of $G{+}V$ over $G$ vary with reasoner behavior. Table 20 reports accuracy, $\mathrm{AUC}(V)$, an easy/hard split by Geo confidence, mean inter-elicitation dispersion, and $\Delta\mathrm{AUC}((G{+}V) - G)$. Two patterns hold. First, settings with higher accuracy tend to have

Table 19: **Performance–access–complexity frontier.** Call counts are per question. "Tok. prob." indicates whether token-level probabilities are required; "Reference" indicates whether a per-question gold answer or trusted verifier is required at inference. AUROC entries are the medians reported elsewhere in the paper and are not all measured on the same panel: $G$, $G{+}V$, $C{+}G{+}V$ and SC are open-ended 18-setting panel medians, while the verbal-only and privileged-likelihood entries are from the aligned GPT-4.1 GPQA comparison of Section 4.6. They should be read as configuration summaries, not as a single controlled ranking.

| **Configuration** | Reasoner | Proposer | Judge | Embed. | Tok. prob. | Reference | AUROC |
|---|---|---|---|---|---|---|---|
| Verbal only ($V$) | $K{+}1$ | — | — | — | no | no | 0.648 |
| Geometry ($G$) | $K$ | — | — | yes | no | no[†] | 0.670 |
| Geometry + verbal ($G{+}V$) | $2K$ | — | — | yes | no | no[†] | 0.691 |
| Reference-assisted $C{+}G{+}V$ | $2K$ | $K_P$ | $K_P$ | yes | no | **yes** | **0.780** |
| Self-consistency SC@8 | 8 | — | — | — | no | no | 0.705 |
| Privileged likelihood (mean logprob) | 1 | — | — | — | **yes** | no | 0.658 |

[†] Requires candidate anchors at inference, but no per-question gold or gold-equivalence signal.

stronger $V$ (Spearman $\rho{=}0.71$, $p{=}0.11$ over six combinations). Second, the incremental value of $V$ is largest when verbalizations are internally stable and the reasoner is already competent: the biggest lift occurs on MMLU-Pro / Claude ($+0.112$ AUC) with low dispersion ($0.021$), whereas GPQA / Claude combines the highest dispersion ($0.127$) with almost no lift ($+0.006$). MedQA / Gemini is the useful counterexample: standalone $V$ is strong ($0.751$) but adds almost nothing over $G$ ($+0.002$), showing that the relevant question is conditional information, not standalone calibration.

Table 20: Verbalization sensitivity audit on the six closed-set settings. The easy/hard split uses Geo confidence as a proxy for question difficulty. $V_{\text{disp}}$ is the mean within-question standard deviation of the generator's numeric self-ratings over repeated verbalization calls.

| Combination | acc | AUC($V$) | AUC($V \mid easy$) | AUC($V \mid hard$) | $V_{\text{disp}}$ | $\Delta((G{+}V) - G)$ |
|---|---|---|---|---|---|---|
| MedQA / Gemini | 0.805 | 0.751 | 0.389 | 0.727 | 0.064 | $+0.002$ |
| MedQA / Claude | 0.744 | 0.787 | 0.769 | 0.719 | 0.055 | $+0.017$ |
| GPQA / Gemini | 0.653 | 0.628 | 0.665 | 0.579 | 0.074 | $+0.055$ |
| GPQA / Claude | 0.508 | 0.580 | 0.558 | 0.565 | 0.127 | $+0.006$ |
| MMLU-Pro / Gemini | 0.800 | 0.633 | 0.697 | 0.622 | 0.011 | $+0.016$ |
| MMLU-Pro / Claude | 0.725 | 0.738 | 0.898 | 0.694 | 0.021 | $+0.112$ |

### E.6 Jargon-free control

The strongest version of the "embedder prior" concern is lexical rather than architectural: perhaps the geometric channel is simply matching medical jargon. We therefore rerun the MedQA audit under two masking schemes. The first applies a global hash-based codebook to content words; this removes human-readable jargon but preserves a question-independent word identity map. The second is the stricter control: a per-question randomized codebook that redraws the word-to-token map independently for every question, so the same lexical item cannot help across questions. Table 21 shows the outcome. On MedQA / Gemini, the per-question codebook still leaves Geo@4 above SC@8 with $\Delta\text{AUC} = +0.065$ and one-sided $p{=}0.031$, despite the drop from $0.848$ to $0.673$ in raw Geo AUC. On MedQA / Claude, the per-question point estimate also remains positive ($+0.038$) but is uncertain at this sample size. The right conclusion is therefore not lexical invariance. It is a lower bound on the structural component: removing cross-question lexical reuse weakens the signal, but does not eliminate it on the settings tested.

Table 21: Jargon-free lexical controls on the two MedQA settings that were re-embedded under masking. The per-question codebook is the stricter control because it destroys cross-question lexical identity.

| Combination | Geo@4 (unmasked) | Geo@4 (global hash) | Geo@4 (per-q codebook) | $\Delta(\text{GEO@4} - \text{SC@8})$ per-q |
|---|---|---|---|---|
| MedQA / Gemini | 0.848 | 0.642 | 0.673 | $+0.065$ |
| MedQA / Claude | 0.760 | 0.727 | 0.666 | $+0.038$ |

Table 22: **Direct trajectory-as-question diagnostic.** AUC for retrieving the gold answer using either the pooled raw-question embedding or the averaged trajectory tail (last 2,000 characters, mean of unit-normalized sample tails). "Hard" uses semantically similar negatives from nearest-neighbor questions; "NN-correct" uses the nearest-neighbor question's correct answer itself. GPQA reports Gemini only because that is the cache currently available for this diagnostic.

| **Benchmark** | **Model** | $N$ | **Hard AUC** raw $\rightarrow$ traj | **NN-correct AUC** raw $\rightarrow$ traj |
|---|---|---|---|---|
| MedQA train | Gemini 3.1 Pro | 1000 | $0.710 \rightarrow 0.873$ | $0.632 \rightarrow 0.806$ |
| MedQA train | Claude Sonnet 4.6 | 1000 | $0.710 \rightarrow 0.856$ | $0.632 \rightarrow 0.785$ |
| MedQA heldout | Gemini 3.1 Pro | 273 | $0.798 \rightarrow 0.924$ | $0.726 \rightarrow 0.867$ |
| MedQA heldout | Claude Sonnet 4.6 | 273 | $0.798 \rightarrow 0.903$ | $0.726 \rightarrow 0.846$ |
| GPQA Diamond | Gemini 3.1 Pro | 198 | $0.665 \rightarrow 0.797$ | $0.626 \rightarrow 0.742$ |
| MMLU-Pro | Gemini 3.1 Pro | 200 | $0.825 \rightarrow 0.920$ | $0.775 \rightarrow 0.874$ |
| MMLU-Pro | Claude Sonnet 4.6 | 200 | $0.825 \rightarrow 0.907$ | $0.775 \rightarrow 0.856$ |

# F Trajectory-as-Question Diagnostic

For completeness we report the direct trajectory-tail retrieval diagnostic noted briefly in Section 4 in table form. The test compares the pooled raw-question embedding to the averaged trajectory tail on identical candidate-answer sets. In every available setting the trajectory tail is the stronger answer-bearing representation.

# G Auxiliary Feature Definitions

Let $e^{(k)}$ denote the condensed final answer extracted from $\mathcal{C}^{(k)}$ by the zero-shot prompt of Appendix H, and let $\phi$ be the embedding model of Section 3.1. The four auxiliary features of Section 3.2 are:

$$\textsc{Str} \;=\; \tfrac{1}{K}\max_{s}\,|\{k : e^{(k)} = s\}|, \tag{7}$$

$$\textsc{Sem} \;=\; \binom{K}{2}^{-1}\sum_{i<j}\frac{\langle \phi(e^{(i)}),\, \phi(e^{(j)})\rangle}{\|\phi(e^{(i)})\|\,\|\phi(e^{(j)})\|}, \tag{8}$$

$$\textsc{Vol} \;=\; \tfrac{1}{K}\sum_{k}\tfrac{1}{M_k-1}\sum_{m=1}^{M_k-1} d\big(\mathbf{z}_{m-1}^{(k)},\, \mathbf{z}_{m}^{(k)}\big), \tag{9}$$

$$\textsc{Ens} \;=\; \mathbb{1}\Big[\mathrm{maj}_k\, e^{(k)}_{\text{Gemini}} = \mathrm{maj}_k\, e^{(k)}_{\text{Sonnet}}\Big]. \tag{10}$$

All features are standardized to zero mean and unit variance on the training fold before being passed to the composite logistic regression; for standalone AUC we use $-\textsc{Vol}$, since higher volatility indicates a more meandering trajectory.

# H Prompt Template

The following prompt template is used to elicit zero-shot chain-of-thought reasoning without exposing the candidate options to the model:

```
You are an expert taking a highly advanced examination.
Please answer the following question.

Question: {question_text}

Do not provide the final answer immediately.
First, think step by step. Explain your logical
reasoning, referencing specific domain knowledge,
mechanisms, and principles.

Provide your reasoning in a clear, step-by-step format.
```

**Prompt V1: original verbalized-confidence elicitation.** This is the prompt used for the $V$ channel in the three-channel decomposition of Section 5 and in every $C+G+V$ result. The model is shown the question and its own committed reasoning trace, and asked for a single integer.

```
You previously reasoned about the following question without seeing any
answer options.

Question:
{question_text}

Your reasoning (which contains your committed answer):
{cot}

On a scale of 0 to 100, how confident are you that the committed answer
in your reasoning is correct? 0 means you are certain it is wrong; 100
means you are certain it is correct. Reply with ONLY a single integer
between 0 and 100 inclusive, no other text.
```

**Prompt V2: stronger post-hoc elicitation adapted from Yoon et al. [2025].** This second prompt is used only for the strong-baseline comparison of Section 4.6 (the aligned GPT-4.1 GPQA panel, $N{=}194$), where it is the entry reported as "adapted Yoon-style verbal confidence" with AUROC $0.648$. It is *adapted from*, not identical to, the protocol of Yoon et al. [2025]: we retain their three-step structure (solution reasoning, confidence reasoning, and confidence verbalization into one of ten labelled bins) and their post-hoc placement relative to a completed reasoning trace, but our elicitation is single-turn on a cached CoT rather than issued as a fresh generation, and we linearly map each reported bin to its midpoint probability to obtain a continuous AUROC-comparable score.

```
First, reason through the question step by step to arrive at an answer.
Then, thoroughly assess your confidence in that answer by evaluating
your thinking process so far.
Finally, classify your confidence into one of the following classes
based on how likely your answer is to be correct:
- "Almost no chance" (0.0-0.1)
- "Highly unlikely" (0.1-0.2)
- "Chances are slight" (0.2-0.3)
- "Unlikely" (0.3-0.4)
- "Less than even" (0.4-0.5)
- "Better than even" (0.5-0.6)
- "Likely" (0.6-0.7)
- "Very good chance" (0.7-0.8)
- "Highly likely" (0.8-0.9)
- "Almost certain" (0.9-1.0)
Each category reflects the probability that your answer is correct.

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

This is the verbatim inference prompt of Yoon et al. [2025] (their Appendix B.2, Listing 1), reused unmodified for the post-hoc single-turn elicitation on our cached traces; only the surrounding single-turn/post-hoc application, not the prompt wording itself, differs from their original multi-dataset generation protocol.

All results labelled "verbal" or "$V$" in the three-channel sections use Prompt V1. All results labelled "adapted Yoon-style" in Section 4.6 use Prompt V2. No result in this paper mixes the two elicitations within a single comparison.

# I Bayesian classification

We will consider the probabilistic use of the CoT-geometry from a bayesian perspective. Let $J$ be the number of answers, and let $h_i$ be a one-hot encoding vector of dimension $J$, where the

$j$'th component, $(h_i)_j$, represents the correctness of answer $a_j$, given that the right answer is $a_i$, ie. $(h_i)_j = \delta_{ij}$. For the general case of open-ended questions we shall also allow for a 0-state, $h_0 = (0, \cdots, 0)$, representing the case where none of the answers are correct. Let $\mathcal{T}$ denote the reasoning trace and let $P(\mathcal{T}|a_i = 1)$ and $P(\mathcal{T}|a_i = 0)$ denote the probabilities of observing $\mathcal{T}$ given that the answer, $a_i$, is respectively correct and incorrect. The likelihood for $h_0$ then reads

$$P(\mathcal{T}|h_0) = \prod_{j\geq 1} P(\mathcal{T}|a_j = 0) \tag{11}$$

and for $i \geq 1$

$$P(\mathcal{T}|h_i) = P(\mathcal{T}|a_i = 1)\prod_{j\neq i} P(\mathcal{T}|a_j = 0) = \frac{P(\mathcal{T}|a_i = 1)}{P(\mathcal{T}|a_i = 0)} P(\mathcal{T}|h_0). \tag{12}$$

Let $\omega_i(\mathcal{T}) = \frac{P(\mathcal{T}|a_i=1)}{P(\mathcal{T}|a_i=0)}$ denote the likelihood ratios for $i \geq 1$ and let $P(h_i)$ denote the prior probabilities over the latent variables $h_0, \cdots, h_J$. Then

$$P(\mathcal{T}) = \sum_{i\geq 0} P(\mathcal{T}|h_i)P(h_i) = P(\mathcal{T}|h_0)\left(\sum_{i\geq 0} \omega_i(\mathcal{T})P(h_i)\right), \tag{13}$$

where we have defined $\omega_0(\mathcal{T}) = 1$. Thus, the posteriors become

$$P(h_i|\mathcal{T}) = \frac{P(\mathcal{T}|h_i)P(h_i)}{P(\mathcal{T})} = \frac{\omega_i(\mathcal{T})P(h_i)}{\sum_{j\geq 0}\omega_j(\mathcal{T})P(h_j)}. \tag{14}$$

We simply set $P(h_0) = 0$ for the case where we know the correct answer is in the proposal set $\{a_i\}_{i=1}^J$.

If we now parameterize the log-likelihood ratio by a scalar trajectory score, $\log \omega_i(\mathcal{T}) = \beta s_i(\mathcal{T})$, then the posterior becomes

$$P(h_i|\mathcal{T}) = \frac{\exp\big(\beta s_i(\mathcal{T}) + \log P(h_i)\big)}{\sum_{j\geq 1}\exp\big(\beta s_j(\mathcal{T}) + \log P(h_j)\big)}. \tag{15}$$

Under uniform priors over the proposal set this reduces to a softmax over $s_i(\mathcal{T})$. Taking $s_i(\mathcal{T}) = -D_i(\mathcal{T})$, with $D_i$ the trajectory-to-option cosine distance from Section 3.2, recovers exactly the softmax form used for GEO. In this sense, our empirical Geo score is a heuristic black-box approximation of the Bayesian posterior above, with trajectory-to-option similarity serving as a proxy for the per-option log-likelihood ratio. The mapping is motivated by the parameterization $\log \omega_i(\mathcal{T}) = -\beta D_i(\mathcal{T})$ rather than validated as the true generative likelihood; the falsification check we run (Table 23) tests only that the all-option and binary projections recover a common $\beta$, not that LLM responses are actually distributed according to $P(\mathcal{T} \mid a_i)$.

The same posterior also admits a *binary projection*: rather than asking for the full vector $\{P(h_i \mid \mathcal{T})\}_{i\geq 1}$, ask only whether the top-scoring option is correct. Under the parameterization $\log \omega_i(\mathcal{T}) = \beta s_i(\mathcal{T})$ and uniform priors, the log-odds of the argmax option against its strongest competitor is

$$\log \frac{P(h_{(1)} \mid \mathcal{T})}{P(h_{(2)} \mid \mathcal{T})} \;=\; \beta\big(s_{(1)}(\mathcal{T}) - s_{(2)}(\mathcal{T})\big), \tag{16}$$

where $s_{(1)} \geq s_{(2)} \geq \cdots$ are the order statistics of the per-option scores; under $s_i(\mathcal{T}) = -D_i(\mathcal{T})$ this score-space margin equals the distance-space top-two margin used in the main paper, $s_{(1)}(\mathcal{T}) - s_{(2)}(\mathcal{T}) = D_{(2)}(\mathcal{T}) - D_{(1)}(\mathcal{T})$, with $D_{(1)} \leq D_{(2)} \leq \cdots$ the sorted per-option distances. The binary correctness probability therefore takes the Fermi–Dirac form $P(\text{argmax correct} \mid \mathcal{T}) \approx \sigma\big(\beta\,(D_{(2)}(\mathcal{T}) - D_{(1)}(\mathcal{T})) - \alpha\big)$, where $\sigma(t) = 1/(1 + e^{-t})$ is the logistic sigmoid and $\alpha$ absorbs both an empirical correctness-rate offset and the contribution of non-runner-up options. With the question-level null-standardization $\tilde{D} = (D_{(2)} - D_{(1)} - \mu_q)/\sigma_q$ of Section 3.2 this becomes $P \approx \sigma(\hat{\beta}_{\text{bin}}\,\tilde{D} - \alpha)$, exactly Eq. 3 of the main text and the score used in the alternative-anchor robustness evaluation of Section 4.5. The all-option GEO score and the Fermi–Dirac score on $\tilde{D}$ are therefore not competing methods but two readouts of the same posterior: the all-option projection (softmax over $J$ scores) and the binary projection on the top-two log-ratio.

### I.1 Empirical audit of the rejection state and the geometric ceiling

We instantiate the formalism above with a two-feature parameterization $\log \omega_i(\mathcal{T}) = \beta_s\, \tilde{s}_i(\mathcal{T}) + \beta_v\, v_i(\mathcal{T})$, where $\tilde{s}_i$ is the null-normalized cosine similarity from question-specific cross-question null trajectories and $v_i$ is the vote share of the $K$ blinded CoTs over option $i$. Parameters $\theta \equiv (\beta_s, \beta_v, \alpha_{h_0})$, with $\alpha_{h_0}$ a scalar log-odds offset on the rejection state $h_0$, are fit by joint MLE on closed-set positives plus null trajectories as $h_0$ exemplars (the discriminative softmax objective); we pool $N{=}198{+}200{+}200$ questions across GPQA Diamond, MedQA-USMLE-4, and MMLU-Pro. To prevent any leakage through either parameter sharing or null-pool sharing, we use 5-fold cross-validation in which both $\theta$ and the per-question null pool are restricted to the training fold; bootstrap confidence intervals use 1000 resamples. This audit uses the cached open-ended anchor-rationale subset rather than the extraction-filtered main-paper split, because the question here is the structural ceiling of the geometric channel rather than direct comparison to Table 1.

**The Bayesian posterior is a tie with top-2 margin under strict CV.** Throughout this audit, $h_{\text{plur}} \equiv h_{i^\star(\mathcal{T})}$ with $i^\star(\mathcal{T}) = \arg\max_i v_i(\mathcal{T})$ denotes the latent "correct" indicator evaluated at the plurality-voted option, i.e. $P(h_{\text{plur}} \mid \mathcal{T})$ is the Bayesian posterior probability that the option the model voted for most is the correct one. Under the leak-free protocol the closed-set correctness AUC of $P(h_{\text{plur}} \mid \mathcal{T})$ is $0.70/0.78/0.73$ on GPQA/MedQA/MMLU-Pro, versus $0.68/0.78/0.69$ for the raw top-2 vote-share margin. The Bayesian posterior is directionally better but not statistically distinguishable from the trivial baseline at this dataset size; the mechanism is that with $K{=}5$ the margin is a six-state discrete score whose ranking AUC ceiling is already close to the continuous Bayesian posterior's. The principled value of the formalism is therefore not better discrimination, but a calibrated, continuous score that the discrete margin cannot supply.

**Cross-projection $\hat{\beta}$ consistency: a falsifiable test of the unification.** The two projections of the unified posterior predict the *same* likelihood-ratio parameter $\beta$. We test this directly by fitting (a) the all-option $\hat{\beta}_{\text{all}}$ on the raw $D[q, j]$ matrix; (b) the binary Fermi–Dirac on the raw top-2 margin $m_q = D_{(2)} - D_{(1)}$, which we call $\hat{\beta}^{\text{raw}}_{\text{bin}}$, predicting $\mathbf{1}[\arg\min_j D_{q,j} = \text{gold}_q]$, with bootstrap CIs (200 resamples); and (c) the binary Fermi–Dirac on the null-standardized margin $\tilde{D} = (m_q - \mu_q)/\sigma_q$ of Section 3.2, which is $\hat{\beta}_{\text{bin}}$ of Eq. 3 in the main text. Because $\beta_{\text{all}}$ multiplies $D$ (gold has *low* $D$, so $\beta_{\text{all}}{<}0$) and $\beta^{\text{raw}}_{\text{bin}}$ multiplies the positive margin $m_q$ (correctness goes up with $m_q$, so $\beta^{\text{raw}}_{\text{bin}}{>}0$), the unified-posterior predictions are $|\hat{\beta}_{\text{all}}| = \hat{\beta}^{\text{raw}}_{\text{bin}} = \hat{\beta}_{\text{bin}}/\bar{\sigma}_q$ (the latter is Eq. 4). Across all six (dataset, model) combinations the bootstrap CIs of $|\hat{\beta}_{\text{all}}|$, $\hat{\beta}^{\text{raw}}_{\text{bin}}$, and $\hat{\beta}_{\text{bin}}/\bar{\sigma}_q$ overlap pairwise (Table 23); the unification therefore survives a falsifiable equality test.

Table 23: Falsification test of the unification (Section 3.2): the all-option softmax and the binary Fermi–Dirac fits are projections of the same posterior with the same $\beta$. Sign-flipped because $\beta_{\text{all}}$ multiplies $D$ (gold has low $D$, $\beta_{\text{all}}{<}0$) and $\beta^{\text{raw}}_{\text{bin}}$ multiplies the positive margin $m_q = D_{(2)} - D_{(1)}$. CIs are 95% over 200 bootstrap resamples of questions. Across all six combinations the three CIs overlap.

| | $\lvert\hat{\beta}_{\text{all}}\rvert$ (raw $D$) | $\hat{\beta}^{\text{raw}}_{\text{bin}}$ (raw $m_q$) | $\hat{\beta}_{\text{bin}}/\bar{\sigma}_q$ (Eq. 4) |
|---|---|---|---|
| GPQA / Gemini | 35.8 [25.6, 50.8] | 29.1 [14.3, 50.9] | 26.2 [ 7.4, 45.6] |
| GPQA / Claude | 38.4 [28.2, 52.3] | 25.0 [ 9.9, 55.8] | 20.7 [ 1.7, 43.2] |
| MedQA / Gemini | 34.1 [29.6, 41.8] | 53.7 [41.1, 80.6] | 33.0 [22.5, 45.6] |
| MedQA / Claude | 33.3 [27.5, 41.1] | 46.5 [34.5, 65.7] | 34.3 [23.8, 47.4] |
| MMLU-Pro / Gemini | 25.7 [20.6, 32.5] | 25.0 [13.5, 38.6] | 23.5 [12.1, 37.8] |
| MMLU-Pro / Claude | 32.6 [26.0, 40.9] | 41.6 [25.9, 63.2] | 40.1 [23.7, 60.3] |

**The rejection prior $P(h_0)$ is rank-invariant.** A common intuition is that abstention can be tuned by setting $P(h_0) = \hat{\pi}_0$ from a held-out estimate of the open-set base rate. Sweeping $\hat{\pi}_0 \in \{10^{-3}, \ldots, 0.99\}$ with $\theta$ fixed leaves the within-closed wrongness AUC *unchanged to four decimals* on all three datasets ($\Delta\text{AUC} < 10^{-4}$). The reason is structural: $\hat{\pi}_0$ multiplies $P(h_0 \mid \mathcal{T})$ by a question-independent factor in the odds ratio, a strictly monotone transform of the score that preserves AUC by construction. The prior moves ROC operating points but cannot change the ranking quality.

**Temperature on $\omega$ plateaus at the max-logit baseline.** A second calibration knob is to rescale the option likelihoods, $\omega_i \to \omega_i^\gamma$. Unlike $\hat{\pi}_0$, this is not rank-invariant: $\gamma \to 0$ collapses options to uniform; $\gamma \to \infty$ gives $P(h_0 \mid \mathcal{T}) \to (1 + (\omega_{\max}/\omega_0)^\gamma \cdot c)^{-1}$, monotone in $-\max_i \log \omega_i$. Empirically the within-closed wrongness AUC rises monotonically in $\gamma$ and *plateaus* at $\gamma \gtrsim 5$: $0.633/0.679/0.555$ on GPQA/MedQA/MMLU-Pro versus $0.625/0.659/0.532$ at $\gamma$=1. The plateau value is the AUC of the single-feature score $-\max_i \log \omega_i$, i.e. the posterior in this limit collapses to a max-logit baseline.

**Anchor-set hypothesis stratification reveals the ceiling.** Stratifying questions by whether the gold answer was ever proposed in the $K$=5 blinded CoTs yields three buckets: CORRECT (plurality = gold), WRONG-IN-SET (gold proposed by some CoT but not the plurality), and NO-HYPOTHESIS-CORRECT (no CoT proposed gold). On all three datasets the mean confidence is non-monotonic in correctness category: CORRECT highest, WRONG-IN-SET *lowest*, NO-HYPOTHESIS-CORRECT back up near CORRECT. The AUC for separating NO-HYPOTHESIS-CORRECT from WRONG-IN-SET is near or below $0.5$ for every score (margin, $P(h_{\text{plur}} \mid \mathcal{T})$, raw $\bar{s}$, $\max_i s_i$) on every dataset. Concretely, on MedQA the means $(\mu_{\text{COR}}, \mu_{\text{W-IN}}, \mu_{\text{NO-HYP}})$ are $(0.97, 0.33, 0.79)$ for margin and $(0.89, 0.51, 0.80)$ for the Bayesian posterior; bucket (iii) scores like bucket (i) because it is dominated by confidently-wrong CoTs whose trajectory geometry is by construction indistinguishable from confidently-correct CoTs.

**Score-family panel: vote-based scores collapse to chance, the embedding channel partially escapes, verbalized confidence escapes the harder separation.** To make the structural ceiling sharp, we evaluate six confidence scores on both bucket separations across all six (benchmark, model) combinations (Table 24). The scores are: GEO (top-2 cosine-similarity margin), Bayesian $P(h_{\text{plur}} \mid \mathcal{T})$ (softmax with $\beta$=20), self-consistency vote share, normalized cluster-entropy $1 - H(v)/\log J$, verbalized confidence (the same generator's 0–100 self-rating, conditional on its own CoT), and an anchor-cohesion control $-\bar{\rho}$. Three patterns hold across the panel. (a) On the easier separation (III) VS (I), vote-based scores (self-consistency, normalized cluster entropy) are at chance on every setting ($95\%$ bootstrap CIs all include $0.5$), while GEO, the Bayesian top-2 margin, and verbalized confidence retain moderate-to-strong signal (AUC $0.60$–$0.89$). (b) On the harder separation (III) VS (II), the structural-ceiling separation, every embedding-based score and every vote-based score is *anti-correlated* with the bucket label (AUC $0.06$–$0.44$): they assign *higher* confidence to no-path-to-gold reasoning than to wrong-but-gold-proposed reasoning. (c) Verbalized confidence is the only score family that escapes this inversion (AUC $0.45$–$0.62$, near or above chance on every setting). The asymmetry between the two separations identifies what the geometric ceiling actually is: (III) traces are not perfectly indistinguishable from (I) traces, the embedding channel detects a residual signal, but they are systematically conflated with the inverse of the wrong-in-set ranking. The verbalized-confidence column, generated by querying the same model that wrote the CoT for a numeric self-rating conditional on its CoT, supplies the second non-geometric channel that the channel-limit argument predicted would be needed; it is the empirical complement that the embedding channel cannot synthesize from itself.

**Anchor-LLM confound and its scope.** Because anchors are themselves emitted by a second LLM (one short committed rationale per option), a portion of bucket (iii) reflects anchor-LLM failure to differentiate options rather than informed-LLM failure to reason. Quantifying this with the anchor-cloud cohesion $\rho_q = \frac{2}{J(J-1)} \sum_{j<j'} \langle \hat{\phi}(a_j), \hat{\phi}(a_{j'}) \rangle$ where $\hat{\phi}(a) = \phi(a)/\|\phi(a)\|$ is the unit-normalized embedding, we find Pearson correlations of $+0.39/+0.51/+0.69$ between $\rho_q$ quintile and bucket-(iii) frequency on GPQA/MedQA/MMLU-Pro. Plurality accuracy drops from $70/85/65\%$ in the lowest-$\rho$ quintile to $45/85/45\%$ in the highest, isolating an anchor-LLM contribution of roughly 10–25 percentage points. However, the within-bucket $\rho$ deltas are small ($\Delta\rho < 0.01$ between bucket (i) and bucket (iii) on every dataset), so anchor cohesion is a population-level diagnostic, not a per-question filter, and bucket-(iii) frequency remains substantial (10–27%) even in the lowest-$\rho$ quintile where anchors are most distinct. The geometric ceiling is therefore real, but the bucket-(iii) headline rate $(35/19/61)$ should be read as an upper bound that includes some anchor-LLM noise.

**Open-set rejection is trivially saturated.** For completeness we evaluate closed-correct vs open-random questions (every option anchor replaced with a random anchor from a different question). The Bayesian $P(h_0 \mid \mathcal{T})$ achieves rejection AUC $0.95/1.00/0.99$, which exceeds the margin baseline ($0.45/0.59/0.51$, near chance because margin is invariant to anchor identity). However the trivial

Table 24: Bucket-separation AUC across six confidence scores and two separations (bootstrapped point estimates over 1000 resamples). (III) VS (I): *can scores tell "confidently-wrong with no path to gold" apart from "correct"?* (III) VS (II): *can scores tell those traces apart from "wrong but gold was proposed by some sample"?* The latter is the structural-ceiling separation: scores fail by being anti-correlated. Verbalized confidence is the only score family that does not invert the ranking.

| | GEO | Bayes | SC | SemEnt | Verbal | $-\bar{\rho}$ |
|---|---|---|---|---|---|---|
| *Separation* (III) VS (I): detect "no-path-to-gold" against "correct" | | | | | | |
| GPQA / Gemini | 0.61 | 0.61 | 0.51 | 0.51 | 0.66 | 0.54 |
| GPQA / Claude | 0.63 | 0.63 | 0.48 | 0.48 | 0.60 | 0.57 |
| MedQA / Gemini | **0.88** | **0.89** | 0.61 | 0.61 | 0.72 | 0.50 |
| MedQA / Claude | **0.78** | **0.80** | 0.51 | 0.51 | 0.77 | 0.51 |
| MMLU-Pro / Gemini | 0.63 | 0.65 | 0.55 | 0.55 | 0.62 | 0.56 |
| MMLU-Pro / Claude | 0.69 | 0.69 | 0.58 | 0.58 | 0.67 | 0.61 |
| *Separation* (III) VS (II): detect "no-path-to-gold" against "wrong-in-set" | | | | | | |
| GPQA / Gemini | 0.23 | 0.30 | 0.14 | 0.13 | **0.62** | 0.47 |
| GPQA / Claude | 0.39 | 0.44 | 0.08 | 0.08 | 0.46 | 0.69 |
| MedQA / Gemini | 0.22 | 0.36 | 0.13 | 0.13 | 0.45 | 0.39 |
| MedQA / Claude | 0.11 | 0.29 | 0.06 | 0.06 | 0.49 | 0.40 |
| MMLU-Pro / Gemini | 0.23 | 0.29 | 0.17 | 0.18 | **0.56** | 0.46 |
| MMLU-Pro / Claude | 0.33 | 0.35 | 0.20 | 0.21 | **0.56** | 0.44 |

baseline $-\overline{s_j}$, the negative mean cosine similarity between trace and the $J$ option anchors, attains AUC 1.000 on all three datasets, exposing that the open-set task collapses to a one-line scalar threshold and is not a paper-worthy claim.

**Summary.** Across rejection prior, likelihood temperature, and feature augmentation, we cannot lift selective-prediction performance above the Bayesian posterior's strict-CV ceiling, which itself ties the top-2 margin within bootstrap CIs. The bucket-(iii) result identifies the structural obstacle: embedding geometry is well-matched to "which option is on-topic" but blind to "is this on-topic option correct," and no calibration knob within the embedding channel can repair that blindness. We therefore report top-2 margin and GEO (its all-option counterpart) as the empirical signals in the main paper, with the Bayesian formalism above as a principled framework that recovers them as special cases under uniform priors and equal-noise likelihoods.

## J Compute, Broader Impacts, and Asset Licenses

**Compute resources.** All reasoner, proposer, judge, verbalization, and embedding calls were issued through vendor APIs. We used Google Vertex AI for Gemini 3.1 Pro and `gemini-embedding-001`, Anthropic / AWS for Claude Sonnet 4.6, Microsoft Azure OpenAI for the `openai_gpt5_mini` judge, and a hosted inference endpoint for Llama 3.3 70B. No local GPU compute was required. Orchestration, caching, fold-wise logistic and softmax fits, AUC bootstraps, and figure generation ran on a single commodity Apple-silicon workstation ($\leq$ 32 GB RAM, $\leq$ 100 GB scratch storage); end-to-end wall-clock per cell is dominated by API latency rather than local compute. Per-question API spend, which is the operative cost axis in this setting, is reported per (benchmark, reasoner) cell in Appendix E.4 (Table 18). The full research project, including failed pilots and abandoned ablations, used at most $\sim 3\times$ the API spend reported there.

**Broader impacts.** *Positive.* Better-calibrated black-box confidence scores support safer abstention and selective prediction when chain-of-thought reasoners are deployed through text-only APIs in high-stakes settings such as clinical decision support (cf. MedQA), where unwarranted confident-but-wrong outputs are the dominant failure mode; the three-channel decomposition also gives downstream users an auditable handle on *why* confidence is high or low (coverage vs. within-trace commitment vs. self-report). *Negative and misuse.* Better-calibrated scores can also induce automation bias: users may defer to a high-confidence model output even when the underlying reasoning is flawed, and the structural ceiling identified in Section 5.1 (bucket-(iii) traces commit confidently to a wrong in-set hypothesis) is not removed by any within-trace calibration knob. The Coverage channel additionally inherits whatever priors the LLM judge encodes; deployments should not treat $C$ as a model-internal

property and should re-validate it under their own judge. The Limitations paragraph in Section 6 flags judge-mediation, the conditional nature of $V$, and the within-trace ceiling as the operational caveats for downstream users.

**Asset licenses and terms of use.** Benchmark datasets are used under their published licenses and cited to their source papers: MedQA-USMLE Jin et al. [2021], GPQA Diamond Rein et al. [2023], and MMLU-Pro; we use them under the terms posted on each dataset's official release page at the time of access (May 2026), and we redistribute no portion of any of them. External pretrained models are accessed only through their providers' standard APIs and are used subject to those providers' terms of service: Gemini 3.1 Pro and `gemini-embedding-001` via Google Vertex AI, Claude Sonnet 4.6 via Anthropic / AWS, `openai_gpt5_mini` (judge) via Microsoft Azure OpenAI, Llama 3.3 70B under the Llama Community License via a hosted inference endpoint, and the open-weight E5-large-v2 embedder Wang et al. [2022a] (MIT License) run locally. No new datasets, models, or benchmarks are released by this paper.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 list the paper's contributions (the black-box trajectory-confidence score $G$, controlled pairing/position evidence for the signal, and the three-channel $C{+}G{+}V$ decomposition), and these match the empirical results in Sections 3–5 and the headline numbers reported there ($6/6$ closed-set settings, median $\Delta\mathrm{AUC} = +0.075$, fixed-pick control $+0.060$, E5 cross-embedder replication, judge-family swap).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 includes an explicit, numbered *Limitations* paragraph covering Coverage's dependence on a reference or trusted verifier, Geometry's dependence on candidate-anchor adequacy, limited judge-robustness testing, the withdrawn terminal-window mechanistic account, the English-only / commercial-embedder evaluation scope, and the untested extension to agentic actions and tool calls. Section 5.4 separately discusses a judge-confound limitation for $C$.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is primarily empirical and does not present formal theorem-proof results. The mathematical content (Section 3, Appendix I) defines scoring rules and a Bayesian interpretation used as modeling/analysis tools and checked empirically rather than proved as standalone theorems.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sections 3.1, 3.2, 5.2, and 4 specify the datasets, retained $N$, model identities, $K$, $K_P$, $J$, temperature, window/stride, features, cross-validation protocol, and evaluation targets; Appendix H and Appendix G provide the prompt templates and feature definitions. The appendix also records the robustness protocols used for the headline claims (window-position sweep, cross-embedder audit, judge-family swap, and system-cost accounting).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: No code or data artifacts are bundled with this paper. All datasets used (MedQA-USMLE, GPQA Diamond, MMLU-Pro, TadA-Bench where applicable) are publicly available benchmarks accessed under their published terms, and all models are accessed through standard vendor APIs; the paper specifies the exact configuration needed to reproduce results (Sections 3.1, 5.2, Appendix H), but the orchestration code itself is not released alongside this submission.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution.
- The instructions should contain the exact command and environment needed to run to reproduce the results.
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The manuscript reports the train/test protocol and retained sample counts, the blinded/open-ended setups, hyperparameters such as $K$, $T$, $W$, and $S$, and the fitting details for the softmax/logistic models in Sections 3.1, 3.2, and 5.2. Appendix H and Appendix G give the prompt templates and the exact auxiliary feature definitions used in the composite models.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports $95\%$ bootstrap confidence intervals in figures and tables, states the number of bootstrap resamples, and uses pooled one-sided Stouffer $z$ and $p$-values for the main across-cell claims. These appear in the main text and appendices, including the window-position sweep, the matched $C{+}G{+}V$ vs. SC comparisons, the judge-family swap, and the Bayesian-audit tables.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated.

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are documented in Appendix J. All reasoner, proposer, judge, verbalization, and embedding calls were issued through the respective vendor APIs, so no local GPU compute was required; orchestration ran on a single commodity workstation ($\leq$ 32 GB RAM, $\leq$ 100 GB scratch storage) with wall-clock dominated by API latency. Per-question API spend is reported in Appendix E.4 (Table 18) for every (benchmark, reasoner) cell.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, the work complies with the NeurIPS Code of Ethics: it uses public benchmarks and model APIs, and involves no crowdsourcing or human-subject experimentation. The paper explicitly discusses technical limitations and the judge-mediated nature of one channel rather than overstating the findings.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: A dedicated broader-impacts paragraph appears in Appendix J. Positive impact: improved black-box confidence estimation supports safer abstention and selective prediction when chain-of-thought reasoners are deployed through text-only APIs in high-stakes settings such as clinical decision support (cf. MedQA). Negative impact and misuse: better-calibrated scores can induce automation bias, the Coverage channel inherits judge-side priors, and the bucket-(iii) failure of Section 5.1 remains structurally unrecoverable from a single trace; the Limitations paragraph in Section 6 flags these caveats for downstream users.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses, fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new frontier language model, image generator, or scraped dataset. It evaluates existing models and benchmarks, so this safeguards item is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Asset provenance, licenses, and API terms-of-service are enumerated in Appendix J: benchmark datasets (MedQA-USMLE, GPQA Diamond, MMLU-Pro) are used under the terms posted on each official release page and cited to their source papers; closed-source models (Gemini, Claude, `openai_gpt5_mini`) are accessed only through their vendors' standard APIs subject to those providers' ToS; open-weight assets (Llama 3.3 70B, E5-large-v2) are used under their published licenses. No new datasets, models, or benchmarks are released by this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new code, model, or dataset assets alongside the camera-ready version.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, annotation labor, or experiments with human participants. All evaluations are run on existing benchmarks and model outputs.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects were recruited or studied, so IRB approval and participant-risk disclosure were not applicable to this work.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are central to the method and are explicitly described throughout the paper: the reasoner, proposer, judge, and verbalization components are all LLM-based, with model identities and roles stated in the main text, and prompt templates provided in Appendix H. This is therefore a core methodological use, not incidental writing assistance. LLMs were additionally used for writing-level edits (copy-editing, phrasing, and proofreading).

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.