EFFICIENT BEST-OF-N WITH RADIAL CONSENSUS SCORE

Anonymous authors
Paper under double-blind review

ABSTRACT

Solving complex problems with large language models (LLMs) often requires selecting the best answer from multiple candidate responses (best-of-N), yet identifying the most reliable one remains challenging, particularly when correctness does not align with majority agreement. Existing approaches based on voting or probability scoring rely on surface-level signals (e.g., frequency or token-level likelihood), often missing high-quality rare answers and ignoring the geometric structure of candidate representations, where correctness may emerge as semantic consistency rather than surface agreement. To address these limitations, we introduce **Radial Consensus Score (RCS)**, a simple, efficient, and training-free method for best-of-N selection. RCS models semantic consensus by computing a weighted Fréchet mean (semantic center) of answer embeddings and ranking candidates by their radial distance to this center. This enables it to recover semantically coherent correct answers despite lexical fragmentation, even when they are not supported by the majority. Importantly, RCS offers a flexible framework that supports multiple weighting schemes, enabling the integration of arbitrary signals, such as semantic agreement or model confidence, with the position of the center adapting accordingly. Extensive experiments across five open-weight models and various benchmarks demonstrate that our metrics consistently outperform strong baselines by 2–7%, with advantages growing as N increases. Moreover, RCS serves as an effective drop-in replacement for majority voting in multi-agent debate and remains effective in black-box scenarios. These results suggest that geometric consensus can serve as a scalable and reliable alternative to voting-based methods for answer selection.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable progress across diverse tasks, yet they remain prone to generating inconsistent or incorrect outputs, particularly in open-ended or reasoning-intensive scenarios (Wang et al., 2023; Huang et al., 2025). A common strategy to improve reliability is best-of-N sampling: generate multiple candidate responses from the same prompt and select the highest-quality one according to a scoring mechanism (Cobbe et al., 2021; Wang et al., 2024a; Kang et al., 2026). In this framework, performance critically depends not only on the diversity of sampled candidates but, more importantly, on the effectiveness of the selection criterion.

Existing selection methods can be broadly categorized into two groups. The first relies on *external evaluators*, such as reward models or verifiers, to score each candidate (Cobbe et al., 2021; Lightman et al., 2024). While effective, these approaches introduce additional computational overhead and require extra supervision. The second group uses *intrinsic signals* derived from the sampled responses themselves. The dominant paradigm in this category is self-consistency via majority voting, which assumes that correct answers appear most frequently among samples (Wang et al., 2023). This works well when the answer space is well-defined, but it struggles in settings with semantic ambiguity or when high-quality answers are rare despite being semantically close to the consensus (Chen et al., 2023; Kang et al., 2026). Recent work has also proposed ranked or reasoning-aware extensions of self-consistency, such as Borda-style voting, along with theoretical analyses connecting Best-of-N to mode estimation and KL-constrained optimization (Gui et al., 2024; Wang et al., 2025; Cordero-Encinar & Duncan, 2025). Parallel efforts explore weighted aggregation and hybrid selection

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

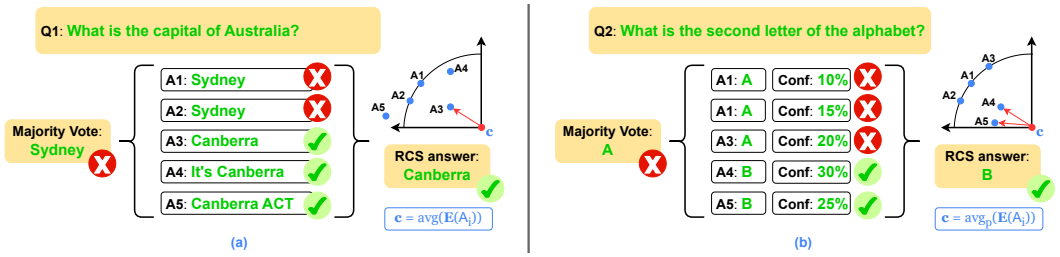

Figure 1: (a) Majority voting ignores semantically similar answers by treating surface forms independently. (b) It also fails to select high-confidence but minority answers. RCS addresses the lexical fragmentation issue by mapping answers into an embedding space via Encoder $\mathbf{E}$ and estimating a centroid $\mathbf{c}$, either *uniformly* or *probability-weighted* over answer embeddings; the latter can also recover high-confidence minority answers. Additional qualitative example is provided in Section 5.3.

strategies, such as correlation-aware multi-agent weighting and best-of-majority, to improve reliability and scaling behavior (Ai et al., 2025; Di et al., 2025).

However, these approaches often rely on additional sampling or surface-form agreement, limiting their practicality in real-world settings, especially when efficient, black-box-compatible, and training-free selection is required. As illustrated in Figure 1, these limitations arise from a mismatch between surface-level signals and the underlying semantic structure of candidate answers. This raises a central question: *Can we design an efficient selection framework that operates directly in the answer space without external supervision or costly sampling?*

To tackle this question, we introduce the **Radial Consensus Score (RCS)**, a simple and effective geometric approach for efficient best-of-N selection. Our key insight is to model semantic consensus explicitly via a weighted Fréchet mean (Fletcher et al., 2004) as the semantic center of answer embeddings and then rank candidates by their radial distance to this center. This formulation naturally incorporates both semantic agreement structure and optional confidence signals (frequency, pairwise cosine similarity, or generation probability) through the choice of weighting distribution P. Different choices of P recover a spectrum of selection strategies, ranging from uniform weighting that treats all candidates equally to frequency-based, cosine-similarity-based, and probability-based schemes that emphasize consensus, mutual agreement, or model confidence. This flexibility allows the framework to adapt to different inference settings while remaining fully training-free and black-box compatible. The resulting score is computationally lightweight, model-agnostic, and directly yields a ranking of individual candidates without requiring clustering or external verifiers. We evaluate RCS on seven established benchmarks spanning short-form QA (SciQ, GPQA), mathematical reasoning (Arithmetics, GSM8K), and long-form multiple-choice tasks (MMLU Formal Logic, MMLU-Pro, BigBenchHard), using five diverse open-weight models. Our metrics consistently outperform strong baselines by 2–7% in selection accuracy, becoming more pronounced as the sampling budget increases. Additional analyses confirm their efficiency and robustness in black-box settings and multi-agent debate scenarios. Our contributions are threefold: **First**, we introduce Radial Consensus Score (RCS) as a geometric consensus method for best-of-N selection that unifies probability- and semantic-based methods into a closed-form estimator without clustering, and provably outperforms voting in both infinite- and finite-sample regimes under mild assumptions. **Second**, comprehensive experiments across diverse tasks and models demonstrate that RCS consistently outperforms strong baselines, with gains increasing as N grows, highlighting favorable scaling behavior. **Third**, ablation studies further confirm the effectiveness and applicability of RCS, showing its robustness in multi-agent debate and fully black-box settings, along with strong practical efficiency, underscoring its impact as a scalable aggregation method for LLM inference.

2 Related Work

Self-Consistency, in its most common form, performs majority voting over N sampled candidates (Wang et al., 2023). Extensions such as universal and fine-grained self-consistency improve robustness but often incur extra LLM calls for evaluation or refinement (Chen et al., 2023; Wang et al., 2024b). Alternatively, *Self-Certainty* directly leverages the model's output probability distribution

to estimate response quality without external reward models, showing strong scaling with N (Kang et al., 2026). Beyond voting-based strategies, recent advances include ranked and reasoning-aware aggregation, e.g. Borda-style voting (Wang et al., 2025; Wan et al., 2025). From a theoretical perspective, Best-of-N has been analyzed under sampling and alignment objectives, linking it to mode estimation and KL-constrained optimization (Gui et al., 2024; Cordero-Encinar & Duncan, 2025).

Geometric consensus itself is not new. Centroid-based MBR (Deguchi et al., 2024) and trajectory-level clustering (Oh & Lee, 2026; Choi & Li, 2026) already exploit embedding geometry, while *Semantic Consensus Weighting* (Knappe et al., 2024) computes pairwise cosine similarities and selects the highest-scoring trajectory group via a discrete aggregation step, thereby inheriting the minority-answer blindness of majority voting. These approaches differ mainly along two axes: *whether embeddings are grouped* and *what is embedded*. On the first, grouping is unnecessary: under a linear (inner-product) utility, selecting the sample nearest the weighted centroid is exactly MBR decoding (Kumar & Byrne, 2004), so a continuous ranking already realizes the consensus objective without any discrete clustering step. On the second, RCS instead operates directly on final-answer embeddings, avoiding the representation collapse and computational overhead that arise when full reasoning trajectories are embedded.

Uncertainty estimation methods offer a complementary lens: probability-based approaches exploit token likelihoods such as predictive entropy (Guerreiro et al., 2023; Manakul et al., 2023; Nguyen et al., 2026a), while semantic entropy methods cluster responses in embedding space and compute entropy over cluster distributions (Kuhn et al., 2023; Farquhar et al., 2024; Lin et al., 2024; Nikitin et al., 2024). Both families, however, either ignore the model's native probability signals or are sensitive to clustering quality, and neither directly yields per-sample candidate scores for efficient ranking.

In contrast, RCS unifies these directions, bridging probabilistic and semantic perspectives via a closed-form weighted Fréchet mean at the *final-answer* level, avoiding embedding-collapse or discrete grouping step while remaining efficient.

3 Methodology

3.1 Problem Statement

Given a prompt x, we sample N independent generations $\{y_1, \ldots, y_N\}$ from a language model using multinomial decoding, where each y_i denotes a full generated response. We then map each response to its final answer $\{a_i\}_{i=1}^N$, where a_i is the extracted or normalized prediction implied by y_i, optionally associated with generation likelihoods $\{p(y_i \mid x)\}_{i=1}^N$. Our goal is to identify the most reliable answer among the candidates. Formally, we consider the problem:

$$a^* = \arg\max_{a_i} \; Q(a_i|x), \tag{1}$$

where $Q(a_i|x)$ is an unknown latent quality score that ranks candidate answers according to their correctness or reliability for prompt x. Since Q is not directly observable at inference time, it must be inferred from the sampled generations themselves.

3.2 Answer Selection via Geometric Consensus

To approximate the latent ranking signal Q, we model agreement among answers in a continuous semantic space. Formally, we encode each final answer into a vector space using a sentence embedding model $\mathbf{E}$, and then apply explicit ℓ_2-normalization to obtain unit-norm vectors $\mathbf{u}_i \in \mathbb{R}^d$:

$$\mathbf{u}_i = \frac{\mathbf{E}(a_i)}{\|\mathbf{E}(a_i)\|_2}, \qquad \|\mathbf{u}_i\|_2 = 1. \tag{2}$$

In the context of best-of-N, sampled answers form a distribution of semantic interpretations of the same query, and a representative solution should reflect their overall agreement. The Fréchet mean (Fletcher et al., 2004) defines such a centroid as the point that minimizes the expected squared distance to all candidate embeddings in the semantic space. Given embeddings $\{\mathbf{u}_i\}_{i=1}^N \subset \mathbb{R}^d$ with a

Table 1: Different instantiations of the distribution P in RCS, where f_i and $\Pr(y_i|\mathbf{x})$ denote the empirical frequency and model probability of candidate y_i, respectively. For the cosine-similarity-weighted variant, $\cos(\mathbf{u}_i, \mathbf{u}_j)$ denotes the cosine similarity between ℓ_2-normalized embeddings.

Variant	**Distribution P**	**Supported LLMs**
Uniform	$p_i = \frac{1}{N}$	Black-box
Frequency-weighted	$p_i = \frac{f_i}{\sum_j f_j}$	Black-box
Cosine-weighted	$p_i \propto \frac{1}{N}\sum_{j=1}^{N} \cos(\mathbf{u}_i, \mathbf{u}_j)$	Black-box
Probability-weighted	$p_i \propto \Pr(y_i\|\mathbf{x})$	White-box

probability distribution $P{=}\{p_i\}_{i=1}^N$, the semantic center is defined as the Fréchet mean under squared Euclidean loss[1]:

$$\mathbf{c}(P) = \arg\min_{\mathbf{z}} \sum_{i=1}^{N} p_i \|\mathbf{u}_i - \mathbf{z}\|_2^2, \tag{3}$$

This admits the closed-form solution:

$$\mathbf{c}(P) = \sum_{i=1}^{N} p_i \mathbf{u}_i. \tag{4}$$

We adopt the squared Euclidean loss for its simple closed-form solution. When the center is restricted to the discrete set $\{\mathbf{u}_i\}_{i=1}^N$, the problem reduces to selecting a representative sample:

$$\mathbf{c}_{\text{medoid}}(P) = \arg\min_{\mathbf{u}_j} \sum_{i=1}^{N} p_i \|\mathbf{u}_i - \mathbf{u}_j\|_2^2, \tag{5}$$

which corresponds to a weighted medoid. This can be viewed as a discrete counterpart of the Fréchet mean, where the solution is constrained to be one of the observed samples.

Different choices of P (Table 1) induce different notions of semantic consensus. In particular, the uniform distribution recovers the standard unweighted center, while frequency-based, similarity-based or likelihood-based distributions incorporate empirical or model-derived reliability signals. Conceptually, $\mathbf{c}(P)$ represents a consensus embedding of the generated answers, where semantically similar answers pull the center toward their shared meaning and low-weight or noisy samples contribute less to the resulting center (see Figure 1).

3.3 Ranking via Radial Consensus Score

Given a semantic center $\mathbf{c}(P)$, we define a ranking function based on the radial dispersion of each candidate embedding relative to this center. For each candidate embedding $\mathbf{u}_i$, its **Radial Consensus Score** (RCS) is defined as:

$$\text{RCS}_i(P) = \|\mathbf{u}_i - \mathbf{c}(P)\|_2. \tag{6}$$

RCS measures how semantically aligned each candidate is with the global consensus induced by the distribution P: smaller values indicate stronger agreement with the collective semantic structure, while larger values suggest divergence or potential noise. We adopt the ℓ_2 norm for its simplicity and consistency with the center formulation (Eq. 3), as the choice of norm does not materially affect the relative ranking of candidates in practice.

The final prediction is selected as the candidate with minimum radial dispersion to the estimated consensus:

$$a^*(P) = \arg\min_{a_i} \text{RCS}_i(P). \tag{7}$$

[1] see Appendix A.1 for details.

When the consensus center is constrained to lie in the discrete candidate set, the final prediction is the candidate corresponding to the weighted medoid:

$$a^*(P) = a_{j^*}, \qquad \mathbf{u}_{j^*} = \mathbf{c}_{\text{medoid}}(P), \tag{8}$$

where $\mathbf{c}_{\text{medoid}}(P)$ is defined in Eq. 5. This selects the candidate whose embedding minimizes the expected *squared* distance to others under P, consistent with the squared-Euclidean objective (Eq. 3). When P is uniform, this reduces to the standard medoid.

Overall, the choice of P controls how consensus is formed: uniform weighting reflects pure geometric agreement, while frequency, cosine similarity, and probability weighting emphasizes stable, semantic agreement and high-confidence responses; all naturally suppressing inconsistent outputs. We refer to the corresponding variants as RCS_{uni}, RCS_{freq}, $\text{RCS}_{\text{cosine}}$, and RCS_{prob}. Notably, the first three depend only on model outputs and embeddings, making them readily applicable in black-box settings. We provide the detailed pseudo-code for our method in Appendix Algorithm 1.

4 Theoretical Analysis

We show that RCS recovers the correct answer whenever it forms the *semantic* majority, even when it is fragmented into several surface forms and therefore holds no plurality under exact-match voting. All statements are on the unit sphere $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$, where ℓ_2-normalized answer embeddings live. Proofs are in Appendix A.2 and A.3.

Generative model. We model the N candidate embeddings as i.i.d. draws from a distribution $\mathcal{D}$ on $\mathbb{S}^{d-1}$ with a well defined mean direction,

$$\mathbf{u}_1, \ldots, \mathbf{u}_N \overset{\text{i.i.d.}}{\sim} \mathcal{D} \text{ on } \mathbb{S}^{d-1}, \qquad \mathbb{E}[\mathbf{u}] = \rho\, \mathbf{c}^*, \quad \rho := \|\mathbb{E}[\mathbf{u}]\|_2 \in (0, 1], \quad \mathbf{c}^* \in \mathbb{S}^{d-1}, \tag{9}$$

where $\mathbf{c}^* = \mathbb{E}[\mathbf{u}]/\|\mathbb{E}[\mathbf{u}]\|_2$ is the true semantic center (a unit vector by construction) and ρ measures concentration. We do *not* assume $\mathcal{D}$ is Gaussian or isotropic in $\mathbb{R}^d$: since $\|\mathbf{u}\|_2 = 1$, a model $\mathbf{u} = \mathbf{c}^* + \boldsymbol{\varepsilon}$ would force $\|\mathbb{E}[\mathbf{u}]\| = \mathbb{E}\|\mathbf{u}\| = 1$, which by Jensen's inequality holds only for $\boldsymbol{\varepsilon} \equiv \mathbf{0}$; Assumption equation 9 avoids this by placing $\mathbb{E}[\mathbf{u}]$ strictly inside the sphere. It holds for the von Mises Fisher law, for any distribution rotationally symmetric about $\mathbf{c}^*$, and for anisotropic $\mathcal{D}$ with nonzero resultant. The weighting $P = \{p_i\}$ (with $\sum_i p_i = 1$) defines the center $\mathbf{c}(P) = \sum_i p_i \mathbf{u}_i$ and covers all four variants of Table 1. Because $\|\mathbf{u}_i\|_2 = 1$, RCS's rule is purely directional,

$$\arg\min_i \|\mathbf{u}_i - \mathbf{c}(P)\|_2 = \arg\max_i \langle \mathbf{u}_i, \mathbf{c}(P) \rangle, \tag{10}$$

so only the direction of $\mathbf{c}(P)$ matters and the analysis is unaffected by the shrinkage ρ.

Proposition 1 (**Asymptotic recovery**). Under Assumption equation 9, for all four weighting variants the center is directionally consistent, $\mathbf{c}(P)/\|\mathbf{c}(P)\|_2 \xrightarrow{a.s.} \mathbf{c}^*$ (for the cosine and probability variants, provided $\mathcal{D}$ is rotationally symmetric about $\mathbf{c}^*$ and, for probability, that the model likelihood increases with alignment to $\mathbf{c}^*$). Consequently $\hat{i} \xrightarrow{a.s.} \arg\max_i \langle \mathbf{u}_i, \mathbf{c}^* \rangle$ and $\Pr(\text{RCS selects the correct answer}) \to 1$ as $N \to \infty$, whereas exact-match voting selects the correct semantic cluster with probability at most $1/L$ when that cluster is split across $L \geq 2$ surface forms with no plurality. Thus RCS is asymptotically superior whenever surface fragmentation occurs.

Proof. See Appendix A.2. □

Proposition 2 (**Finite sample gap**). Let the candidates split into cluster A (correct, size k, mean direction $\boldsymbol{\mu}_A$) and cluster B (spurious, size $N - k$, mean direction $\boldsymbol{\mu}_B$), with common concentration ρ and angular separation $\theta = \angle(\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)$. Define the signed margin $\Delta = \rho \sin(\theta/2)\,(2k - N)/N$. For the geometric group (uniform and frequency), if $\Delta > 0$ then

$$\Pr(\text{RCS selects } A) \;\geq\; 1 - \exp\big(-N\Delta^2/2\big). \tag{11}$$

Uniform weighting gives $\Delta > 0 \iff k > N/2$; frequency weighting replaces $(2k - N)/N$ by the signed weighted mass, recovering a lexically dominant correct cluster even at $k < N/2$. Cosine weighting has no finite sample guarantee, as it amplifies the tightest cluster regardless of correctness; probability weighting can recover a genuine semantic minority when its likelihood aligns with correctness. When $\Delta < 0$, RCS follows the majority and fails as voting does, while asymptotic recovery (Proposition 1) still holds.

Proof. See Appendix A.3. □

Table 2: Best-of-N selection accuracy. **BBF**: black-box friendly. For non-RCS methods, underline marks the best baseline per setting. Among RCS variants, **bold** indicates the best RCS method per column. No baseline significantly outperforms RCS at the column level ($p < 0.05$, paired t-test); a full per-cell (model $\times$ benchmark) McNemar breakdown against all six baselines is in Appendix Table 9. Additional results in Appendix Table 8.

Model	Method	BBF	SciQ	GPQA	Arithmetics	GSM8K	Form.Log.	Average
Qwen2.5-3B	Greedy	✓	59.3	20.9	89.0	65.0	34.9	53.8
	Oracle	✓	80.8±0.7	51.0±2.9	99.5±0.5	93.1±0.8	73.5±5.7	79.6
	NLL	✗	59.4±0.7	15.7±2.4	36.0±0.9	25.0±3.1	23.0±2.7	31.8
	ANLL	✗	59.2±0.5	15.9±2.2	36.0±1.3	24.9±3.3	22.8±3.0	31.8
	SC	✓	64.4±0.6	21.4±0.6	78.7±3.3	52.8±1.3	40.5±2.1	51.6
	SCW	✓	64.5±0.8	23.0±0.3	51.8±3.0	37.4±2.8	25.7±2.4	40.5
	CE	✗	64.0±0.5	16.8±1.7	80.3±4.3	53.3±2.8	41.3±0.8	51.1
	ModeX	✓	62.3±0.1	18.1±2.3	48.5±2.3	33.3±3.1	25.1±2.6	37.5
	$\text{RCS}_{\text{medoid}}$	✓	65.2±1.0	23.8±0.5	81.7±3.0	63.6±2.6	43.6±1.6	55.6
	RCS_{uni}	✓	**65.6±0.4**	23.1±2.2	**85.7±3.2**	**65.8±3.1**	**44.7±1.2**	**57.0**
	RCS_{freq}	✓	64.8±0.7	22.8±0.5	82.8±4.6	59.7±1.6	43.4±1.7	54.7
	$\text{RCS}_{\text{cosine}}$	✓	**65.6±0.6**	**24.2±1.8**	85.2±4.5	65.0±2.8	**44.7±1.2**	56.9
	RCS_{prob}	✗	65.5±0.6	22.5±2.4	84.2±4.0	63.4±1.8	43.9±1.2	55.9
Llama3.2-3B	Greedy	✓	55.6	23.0	96.0	79.7	34.9	58.3
	Oracle	✓	79.8±0.4	52.5±1.5	99.8±0.3	92.6±0.3	71.4±4.4	79.2
	NLL	✗	53.6±1.1	20.2±3.2	86.5±1.3	74.3±1.5	33.3±2.1	53.6
	ANLL	✗	53.6±0.5	19.7±0.5	88.5±2.6	73.9±2.4	29.9±6.0	53.1
	SC	✓	58.8±1.1	17.4±2.6	97.5±1.3	78.0±0.8	16.1±3.2	53.6
	SCW	✓	59.5±1.0	25.4±0.9	69.2±2.0	62.6±1.1	18.5±3.3	47.0
	CE	✗	59.3±0.6	20.9±2.2	97.3±1.3	78.5±0.8	17.5±2.7	54.7
	ModeX	✓	53.2±0.6	18.1±2.1	70.0±2.3	62.6±2.1	16.9±0.5	44.2
	$\text{RCS}_{\text{medoid}}$	✓	**60.0±1.2**	25.6±0.8	**98.3±1.0**	80.5±1.5	19.6±3.7	56.8
	RCS_{uni}	✓	59.5±0.8	25.0±0.8	98.2±1.0	80.5±1.5	20.4±3.2	56.7
	RCS_{freq}	✓	59.8±1.0	25.4±0.9	98.2±1.0	79.4±1.3	18.0±3.7	56.2
	$\text{RCS}_{\text{cosine}}$	✓	59.8±1.2	25.2±1.2	98.0±0.9	81.1±1.6	20.6±3.6	56.9
	RCS_{prob}	✗	59.6±0.8	**26.2±1.3**	97.5±0.8	**84.1±0.8**	**33.1±3.2**	**60.2**

5 Experiment

5.1 Experiment Setup

Models And Datasets. We evaluate on five popular open-weight models from distinct families: Qwen2.5-3B, 7B (Yang et al., 2025), Llama3.2-3B, Llama3.1-8B (Grattafiori et al., 2024), and Gemma2-9B (Riviere et al., 2024). We use seven established benchmarks covering both short-form question-answering and long-form reasoning tasks with different answer formats: (1) Short-form QA: **SciQ** (Welbl et al., 2017) and GPQA Diamond (**GPQA** (Rein et al., 2024)), (2) Long-form Math Reasoning: **Arithmetics** (Choi et al., 2025), **GSM8K** (Cobbe et al., 2021), and (3) Long-form QA multiple choice: MMLU Formal Logic (**Form.Log.** (Hendrycks et al., 2021)), **MMLU-Pro** (Wang et al., 2024c), BigBenchHard (**BBH-Date**, **BBH-Navigate** (Suzgun et al., 2023)).

Baselines. We compare our method against the following baselines, which require only single generation pass: (1) **NLL** (Nguyen et al., 2026a), which selects answers with the highest likelihood based on negative log-likelihood; (2) **ANLL** (Guerreiro et al., 2023), which uses average negative log-likelihood for length-normalized scoring; (3) Self-Consistency (**SC** (Wang et al., 2023)), selecting the most frequent answer; (4) Semantic Consensus Weighting (**SCW** (Knappe et al., 2024)), which embeds full reasoning paths, computes pairwise cosine similarities, groups scores by unique reasoning path, and selects the highest-scoring group; (5) Self-Certainty (**CE** (Kang et al., 2026)), which ranks answers by combining frequency and model confidence using Borda voting; and (6) **ModeX** (Choi & Li, 2026), which applies spectral clustering over reasoning trajectories and selects answers from the resulting cluster centroids. We report four RCS variants (Table 1): $\textbf{RCS}_{\textbf{uni}}$, $\textbf{RCS}_{\textbf{freq}}$, $\textbf{RCS}_{\textbf{cosine}}$, and $\textbf{RCS}_{\textbf{prob}}$ (using ANLL as a length-normalized probability proxy). We also report $\textbf{RCS}_{\textbf{medoid}}$ (uniform, unweighted) as a discrete baseline. All RCS variants use `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019). We sample N=10 responses (also N=5, 20, 40) at T=1 via vLLM (Kwon et al., 2023); results are mean±std over three seeds. Additional details are in Appendix A.5.

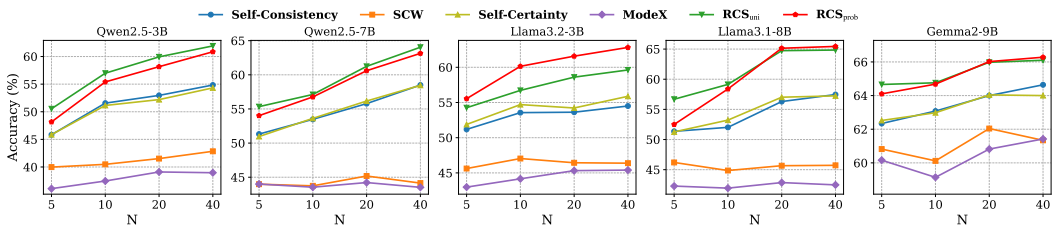

Figure 2: Average performance over five benchmarks for different numbers of sampling responses N. Each subplot corresponds to an LLM backbone, showing how accuracy changes with increasing N.

Table 3: Performance on Arithmetics and Form.Log. with Multi-agent Debate. Best values are bolded. **R=0,1,2** indicate debate rounds, while **Avg.** denotes the average over all rounds.

Model	Method	N	Arithmetics				Form.Log.			
			R=0	R=1	R=2	Avg.	R=0	R=1	R=2	Avg.
Qwen2.5-3B	SC	10	94.5±4.0	60.4±1.6	34.6±5.8	63.1	44.8±2.5	35.3±2.9	27.6±5.4	35.9
	RCS$_{uni}$	10	94.3±3.7	59.5±4.1	34.0±11.0	62.6	**46.0±2.3**	35.8±5.3	**28.8±6.0**	**36.9**
	RCS$_{freq}$	10	**95.2±3.6**	**61.4±1.2**	**34.7±5.4**	**63.8**	44.8±2.8	**36.1±3.3**	28.7±5.5	36.5
	SC	20	96.5±0.5	24.8±3.0	34.0±3.3	51.8	42.3±0.5	40.5±1.6	35.7±1.6	39.5
	RCS$_{uni}$	20	**97.7±1.0**	**27.3±3.2**	**34.5±5.1**	**53.2**	**46.6±0.9**	**45.2±1.4**	**37.8±0.9**	**43.2**
	RCS$_{freq}$	20	97.2±0.3	26.0±2.2	34.3±4.0	52.5	42.6±0.5	42.1±0.8	35.7±1.6	40.1
Llama3.2-3B	SC	10	80.7±2.0	85.6±10.7	75.8±5.9	80.7	31.3±2.9	28.4±7.8	29.5±7.7	29.7
	RCS$_{uni}$	10	79.1±4.3	**85.8±9.1**	**76.2±4.6**	80.4	**33.2±0.5**	**28.7±8.0**	**29.9±6.4**	**30.6**
	RCS$_{freq}$	10	**81.1±12.7**	85.6±10.0	**76.2±5.1**	**81.0**	32.1±2.9	28.6±8.3	29.5±16.6	30.1
	SC	20	99.0±0.5	72.2±4.1	41.5±2.2	70.9	13.2±0.5	3.2±0.8	3.7±1.7	6.7
	RCS$_{uni}$	20	99.0±1.0	**76.5±4.3**	**43.3±1.5**	**72.9**	**21.4±2.4**	**9.0±1.2**	**7.1±2.4**	**12.5**
	RCS$_{freq}$	20	**99.2±0.8**	73.8±3.8	42.5±2.6	71.8	15.6±1.7	4.2±0.5	3.7±0.9	7.8

5.2 Main Results

RCS Consistently Outperforms Baselines. We report selection accuracy in Table 2. RCS variants consistently achieve top performance across settings. RCS$_{uni}$, RCS$_{cosine}$, and RCS$_{prob}$ rank first overall, outperforming the best baseline (CE) by 2–6%. Notably, RCS$_{cosine}$ outperforms SCW despite using the same pairwise cosine signals, highlighting the advantage of operating at the final-answer level and avoiding discrete grouping. Similarly, RCS$_{freq}$ outperforms SC by avoiding hard tie-breaking via soft semantic weighting. Among baselines, CE is the strongest competitor, while ModeX and SCW both underperform due to full-reasoning embedding collapse (see Section 5.3). The gains are model-dependent rather than scale-dependent, with RCS benefiting most when answers exhibit greater lexical diversity. Overall, a unified geometric framework with flexible weighting schemes is more expressive and robust than surface-form or trajectory-level aggregation.

RCS Scales Effectively With Increasing Sampling Budget. Figure 2 shows performance as the number of sampled responses N increases with only top methods are shown. We highlight RCS$_{uni}$ and RCS$_{prob}$, the best-performing variants from Table 2 (full results in Appendix A.6). RCS consistently outperforms all baselines at larger N (e.g., N=20, 40), with gains up to +7%. As N increases, correct answers may remain in the minority; RCS recovers them by anchoring to the semantic center rather than to frequency, maintaining robustness even when correct outputs are sparse.

RCS Enhances Multi-Agent Debate. We evaluate RCS as a drop-in replacement for majority voting in multi-agent debate (Du et al., 2024; Choi et al., 2025) under fully black-box settings, using N=10, 20 agents for R=2 rounds with Qwen2.5-3B and Llama3.2-3B. We report RCS$_{uni}$ and RCS$_{freq}$ (Table 3), which are computationally efficient and do not require access to model internals. Increasing the number of agents to N=20 consistently degrades performance across debate rounds, particularly on Llama3.2-3B, suggesting that naive scaling introduces additional variability that destabilizes reasoning (Nguyen et al., 2026b). In contrast, RCS improves performance by 1–2% on average, with larger gains at N=20, highlighting its robustness when aggregation becomes more challenging.

Table 4: Performance under (a) black-box setting (N=10) and (b) complexity comparison.

(a)

Method	GPQA	MMLU-Pro	BBH-Date	BBH-Navigate	Avg.
SC	13.1	47.6	76.8	39.2	44.2
SCW	12.6	47.6	76.2	39.7	44.0
$\text{RCS}_{\text{medoid}}$	**14.6**	**48.6**	76.4	39.2	44.7
RCS_{uni}	14.1	**48.6**	**78.0**	40.8	**45.4**
RCS_{freq}	13.6	47.6	77.6	**41.2**	45.0
$\text{RCS}_{\text{cosine}}$	13.6	47.6	77.2	40.5	44.7

(b)

Method	Complexity
NLL, ANLL	$\mathcal{O}(NL)$
SC	$\mathcal{O}(N)$
SCW	$\mathcal{O}(N^2d)$
CE	$\mathcal{O}(NL\|V\|)$
ModeX	$\mathcal{O}(N^2L+N^3)$
$\text{RCS}_{\text{medoid/cosine}}$	$\mathcal{O}(N^2d)$
$\text{RCS}_{\text{uni/freq}}$	$\mathcal{O}(Nd)$
RCS_{prob}	$\mathcal{O}(N(d+L))$

Question: A spin-1/2 particle sits in a magnetic field along $+Z$; at t=0 the field is instantaneously switched to $+Y$, same magnitude B. What is the oscillation frequency of $\langle S_z \rangle$?

Ground Truth: γB

A1: `(gamma*B)/(2*pi)`. Let me verify...

A2: `g*B/h`

A3: ω=2μB/$\hbar$

A4: `gamma*B`

A5: `gamma*B`. Wait, let me double-check...

A6: `gamma*B` is the final...

A7: 2μB/`\hbar`. Hmm, actually...

A8: `\gamma B.`

A9: `\gamma B`. So the frequency is...

A10: γB

Self-Consistency fails to clearly distinguish between 10 verbatim-distinct strings (✗).

RCS selects γB (✓), as A4–A6, A8–A10 embed near one center despite lexical fragmentation.

Figure 3: Qualitative example showing RCS recovers correct answers more reliably than SC (GPQA, Gemma2-9B).

RCS Improves Under Fully Black-box Setting. We further evaluate a fully black-box setting using Cohere (*command-a-03-2025*) on GPQA, MMLU-Pro, and BigBenchHard (Table 4a). RCS outperforms traditional Self-Consistency, with RCS_{uni} achieving the best performance, yielding +1% average improvement. This highlights the advantage of RCS in capturing consistency beyond simple frequency-based selection.

5.3 Ablation Study

Complexity Analysis. We report computational complexity in Table 4b, where d, L, and $|V|$ denote embedding dimension, sequence length, and vocabulary size. NLL and ANLL scale as $\mathcal{O}(NL)$ due to token-level scoring, while SC is $\mathcal{O}(N)$ since it only performs frequency counting over final answers. More expressive methods such as CE and ModeX capture richer semantic or structural dependencies but incur higher costs due to full token-level distributions or pairwise interactions. In contrast, RCS variants (e.g., RCS_{uni}) avoid pairwise comparisons by estimating a semantic center and dispersion in $\mathcal{O}(Nd)$, achieving a favorable trade-off between efficiency and semantic expressiveness.

Qualitative Example. We present a qualitative example from GPQA (Figure 3). The correct answer (γB) appears in 6/10 samples but splits into six verbatim-distinct strings across notation and trailing chatbot text, so Self-Consistency sees a full 10-way tie and falls back to a wrong tie-break, while RCS_{uni}'s centroid still lands inside the correct cluster. On a harder boundary case from the same instance (N=7 (A1-A7), correct cluster size k=3 $< N/2$), real embeddings still place RCS_{uni}'s centroid nearest a correct candidate (cosine distance 0.498 vs. 0.609–0.899 for the wrong ones), even though this falls outside Proposition 2's $k > N/2$ construction. This highlights the robustness of RCS in aggregating diverse outputs under noisy sampling.

Effect Of Embedding Models. We compare three sentence transformers of increasing capacity in Appendix Figure 6. Results show minimal variation across models. This stability is especially evident on Arithmetics, where answers are numeric. On Form.Log., we observe a slight drop with `all-roberta-large-v1`, particularly for the probability-based variant, suggesting that higher embedding capacity does not necessarily improve semantic alignment for this task.

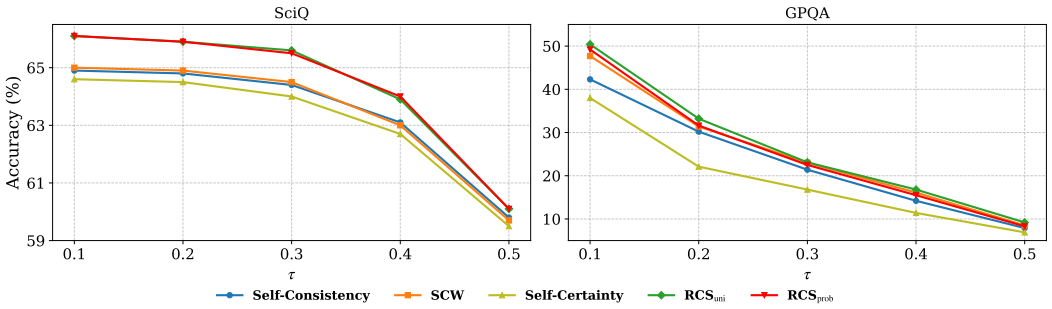

Figure 4: Effect of varying τ using Qwen2.5-3B. Results for Llama3.2-3B are in Appendix Figure 8.

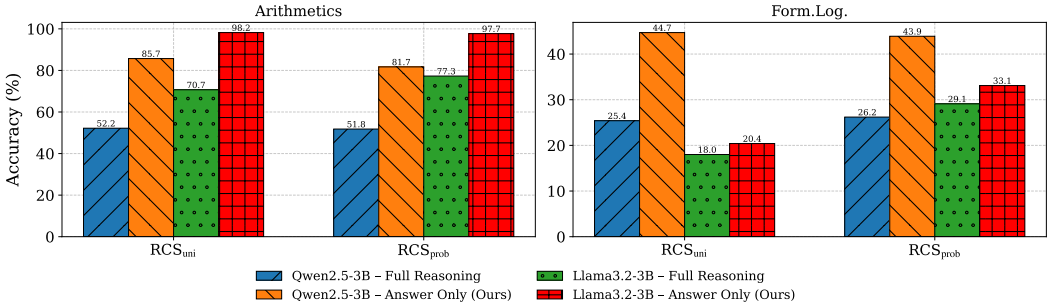

Figure 5: Effect of reasoning paths on Arithmetics and Formal Logic.

Effect Of Correctness Threshold. Figure 4 reports performance on SciQ and GPQA under varying correctness thresholds. We focus on RCS_{uni} and RCS_{prob}, which consistently achieve the best results as the ROUGE-F1 threshold increases from 0.1 to 0.5, particularly within the common range of 0.2–0.4. This demonstrates RCS's robustness to threshold selection.

Effect of Full-Trajectory Embeddings. A natural extension of RCS is to use full generation trajectories instead of final-answer embeddings. However, this leads to an *embedding collapse* effect, where candidates become less distinguishable due to shared prefixes and redundant reasoning. We compare this variant with final-answer embeddings (Figure 5) and observe consistent performance drops. This collapse issue also affects methods that operate on full trajectories, such as ModeX and SCW, which cluster or group complete reasoning traces based on embedding similarity. This suggests that intermediate tokens introduce noise and weaken discriminative signals, while final answers provide a more compact and task-relevant representation. RCS avoids both issues by restricting embeddings to the extracted final answer and ranking all candidates continuously.

6 Conclusion

We introduce Radial Consensus Score (RCS), an efficient method for reliable best-of-N selection in LLMs. RCS ranks candidates by their distance to a weighted Fréchet mean of answer embeddings, where the weighting scheme (uniform, frequency, cosine-similarity, or probability) determines how consensus is defined. Our metrics consistently outperforms strong baselines across diverse benchmarks, with larger gains under higher sampling budgets, while being efficient, black-box compatible, and applicable as a drop-in replacement for majority voting in multi-agent debate. This highlights geometric consensus as a simple and scalable principle for reliable answer selection.

AI use statement

In this work, we used generative AI tools solely to support writing, including language polishing and improving clarity and readability. We did not use generative AI tools to develop research ideas,

design methods, conduct experiments, analyze results, or generate the substantive content of the work. All AI-assisted text was reviewed and revised by the authors, who verified its accuracy and consistency with the work. We take full responsibility for the final content of this work, including all claims and artifacts produced with the aid of generative AI.

Ethics statement

RCS can positively impact applications by improving the reliability of best-of-N selection in LLMs, leading to more consistent and accurate outputs. However, its effectiveness depends on the quality of embedding representations, which may inherit or amplify biases from the underlying model. Therefore, caution is needed when applying it in high-stakes or safety-critical settings.

Reproducibility statement

We provide implementation details in Appendix A.5 and will release the source code openly for research purposes upon publication.

References

Rui Ai, Yuqi Pan, David Simchi-Levi, Milind Tambe, and Haifeng Xu. Beyond majority voting: Llm aggregation by leveraging higher-order information. *arXiv preprint arXiv:2510.01499*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023.

Hyeong Kyu Choi and Sharon Li. Modex: Evaluator-free best-of-n selection for open-ended generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Paula Cordero-Encinar and Andrew B Duncan. Certified self-consistency: Statistical guarantees and test-time training for reliable reasoning in llms. *arXiv preprint arXiv:2510.17472*, 2025.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe, Hideki Tanaka, and Masao Utiyama. Centroid-based efficient minimum Bayes risk decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11009–11018, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.654. URL https://aclanthology.org/2024.findings-acl.654/.

Qiwei Di, Kaixuan Ji, Xuheng Li, Heyang Zhao, and Quanquan Gu. Best-of-majority: Minimax-optimal strategy for pass@ k inference scaling. *arXiv preprint arXiv:2510.03199*, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, 2024.

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

P.T. Fletcher, Conglin Lu, S.M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004. doi: 10.1109/TMI.2004.831793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nuno M Guerreiro, Elena Voita, and André FT Martins. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1059–1075, 2023.

Lin Gui, Cristina Gârbacea, and Victor Veitch. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *Advances in Neural Information Processing Systems*, 37: 2851–2885, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. In *Advances in Neural Information Processing Systems*, volume 38, pp. 19720–19745, 2026.

Tim Knappe, Ryan Li, Ayush Chauhan, Kaylee Chhua, Kevin Zhu, and Sean O'Brien. Semantic self-consistency: Enhancing language model reasoning via semantic weighting. *arXiv preprint arXiv:2410.07839*, 2024.

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

Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/N04-1022/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

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

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024.

Potsawee Manakul, Adian Liusie, and Mark 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*, pp. 9004–9017, 2023.

Manh Nguyen, Sunil Gupta, and Hung Le. Probabilities are all you need: A probability-only approach to uncertainty estimation in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 32546–32554, 2026a.

Manh Nguyen, Anh Nguyen, Dung Nguyen, Svetha Venkatesh, and Hung Le. Hear both sides: Efficient multi-agent debate via diversity-aware message retention. *arXiv preprint arXiv:2603.20640*, 2026b.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Jungsuk Oh and Jay-Yoon Lee. Latent self-consistency for reliable majority-set selection in short-and long-answer reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 32591–32599, 2026.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

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

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient llm sampling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3613–3635, 2025.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Weiqin Wang, Yile Wang, and Hui Huang. Ranked voting based self-consistency of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14410–14426, 2025.

Xinglin Wang, Yiwei Li, Shaoxiong Feng, Peiwen Yuan, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. Integrate the essence and eliminate the dross: Fine-grained self-consistency for free-form language generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11782–11794, 2024b.

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. In *International Conference on Learning Representations*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024c.

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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*, 2017.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A APPENDIX

Algorithm 1 Best-of-N Selection with Radial Consensus Score (RCS)

Require: Answers $\{a_i\}_{i=1}^N$, distribution P, embedding model $\mathbf{E}$, mode $\in \{\text{continuous}, \text{discrete}\}$
Ensure: Selected index i^*
Embeddings: $\mathbf{u}_i = E(a_i)$ for $i = 1, \ldots, N$
if mode = continuous **then**
  $\mathbf{c} \leftarrow \sum_{i=1}^N p_i \mathbf{u}_i$
else
  $\mathbf{c} \leftarrow \arg\min_{\mathbf{u}_j} \sum_{i=1}^N p_i \|\mathbf{u}_i - \mathbf{u}_j\|_2^2$
end if
$i^* \leftarrow \arg\min_i \|\mathbf{u}_i - \mathbf{c}\|_2$
return i^*

A.1 PROOF OF GEOMETRIC CENTER ESTIMATION

Expanding the objective, we have:

$$\sum_{i=1}^N p_i \|\mathbf{u}_i - \mathbf{z}\|_2^2 = \sum_{i=1}^N p_i \left(\|\mathbf{u}_i\|_2^2 - 2\mathbf{u}_i^\top \mathbf{z} + \|\mathbf{z}\|_2^2\right) \tag{12}$$

$$= \sum_{i=1}^N p_i \|\mathbf{u}_i\|_2^2 - 2\mathbf{z}^\top \sum_{i=1}^N p_i \mathbf{u}_i + \|\mathbf{z}\|_2^2. \tag{13}$$

Taking derivative with respect to $\mathbf{z}$ and setting it to zero:

$$-2 \sum_{i=1}^N p_i \mathbf{u}_i + 2\mathbf{z} = 0, \tag{14}$$

which yields:

$$\mathbf{z} = \sum_{i=1}^N p_i \mathbf{u}_i. \tag{15}$$

Since the objective is strictly convex in $\mathbf{z}$, this solution is unique.

A.2 PROOF OF PROPOSITION 1

Setup. The candidates $\mathbf{u}_1, \ldots, \mathbf{u}_N$ are i.i.d. on $\mathbb{S}^{d-1}$ with $\mathbb{E}[\mathbf{u}] = \rho \mathbf{c}^*$ and $\rho > 0$ (Assumption 9). Recall that $\mathbf{c}^* := \mathbb{E}[\mathbf{u}]/\|\mathbb{E}[\mathbf{u}]\|_2$ is a unit vector by construction, so $\mathbf{c}^* \in \mathbb{S}^{d-1}$. Because $\|\mathbf{u}_i\|_2 = 1$, the norm $\|\mathbf{u}\|_2$ is bounded, hence $\mathbb{E}\|\mathbf{u}\|_2 \leq 1 < \infty$; the finite first moment required by the strong law of large numbers (SLLN) holds automatically, with no further integrability assumption. By the selection identity Eq. 10 it suffices to show $\mathbf{c}(P)/\|\mathbf{c}(P)\|_2 \xrightarrow{a.s.} \mathbf{c}^*$, since the ranking $\arg\max_i \langle \mathbf{u}_i, \cdot \rangle$ is continuous in the direction of its argument and invariant to its positive scaling.

Geometric group (uniform, frequency). Here the weights do not depend on $\{\mathbf{u}_i\}$. For uniform weighting $p_i = 1/N$,

$$\mathbf{c}(P) = \frac{1}{N} \sum_{i=1}^N \mathbf{u}_i \xrightarrow{a.s.} \mathbb{E}[\mathbf{u}] = \rho \mathbf{c}^* \tag{16}$$

by the SLLN, so $\mathbf{c}(P)/\|\mathbf{c}(P)\|_2 \to \rho \mathbf{c}^*/\rho = \mathbf{c}^*$ almost surely. For frequency weighting, $p_i = f_i / \sum_j f_j$ with f_i the lexical count of answer i. Writing $\mathbf{c}(P) = \left(\frac{1}{N}\sum_i f_i \mathbf{u}_i\right) / \left(\frac{1}{N}\sum_i f_i\right)$, the counts f_i depend only on the discrete answers, not on the embedding noise given the answers; numerator and denominator each converge a.s. by the SLLN to $\mathbb{E}[f\mathbf{u}]$ and $\mathbb{E}[f] > 0$. Since counts are invariant across paraphrases of the same semantic answer, $\mathbb{E}[f\mathbf{u}]$ is a positive multiple of $\mathbf{c}^*$, and the ratio is directionally consistent.

Endogenous variant (cosine) and exogenous variant (probability). Let $\bar{\mathbf{u}} := \frac{1}{N}\sum_{j=1}^{N}\mathbf{u}_j$ denote the empirical mean embedding. Both variants take the self-normalized form

$$\mathbf{c}(P) = \frac{\frac{1}{N}\sum_i w(\mathbf{u}_i)\,\mathbf{u}_i}{\frac{1}{N}\sum_i w(\mathbf{u}_i)}, \tag{17}$$

where the weight is $w(\mathbf{u}) = \langle \mathbf{u}, \bar{\mathbf{u}}\rangle$ for the cosine variant (the average cosine similarity of $\mathbf{u}$ to the other candidates, as in Table 1), and $w(\mathbf{u}) = \Pr(y|x)$ for the probability variant. Both weights are bounded and continuous on the compact sphere. By the SLLN the numerator converges a.s. to $\mathbb{E}[w(\mathbf{u})\mathbf{u}]$ and the denominator to $\mathbb{E}[w(\mathbf{u})] > 0$; by the continuous mapping theorem,

$$\mathbf{c}(P) \xrightarrow{a.s.} \mathbf{m} := \frac{\mathbb{E}[w(\mathbf{u})\mathbf{u}]}{\mathbb{E}[w(\mathbf{u})]}. \tag{18}$$

(For the cosine variant, $\bar{\mathbf{u}} \xrightarrow{a.s.} \rho\mathbf{c}^*$ by the geometric case above, so the effective weight is $w(\mathbf{u}) \to \rho\langle \mathbf{u}, \mathbf{c}^*\rangle$, a function of $\langle \mathbf{u}, \mathbf{c}^*\rangle$ alone.)

It remains to show $\mathbf{m}$ points along $\mathbf{c}^*$. Assume $\mathcal{D}$ is *rotationally symmetric about* $\mathbf{c}^*$: its density depends on $\mathbf{u}$ only through $t := \langle \mathbf{u}, \mathbf{c}^*\rangle$. Every point on the sphere decomposes exactly as

$$\mathbf{u} = t\,\mathbf{c}^* + \sqrt{1-t^2}\,\mathbf{s}, \qquad t = \langle \mathbf{u}, \mathbf{c}^*\rangle \in [-1,1], \quad \mathbf{s} \in \mathbb{S}^{d-1} \cap (\mathbf{c}^*)^{\perp}, \tag{19}$$

which is an identity for any $\mathbf{u}$, not an assumption. Under rotational symmetry about $\mathbf{c}^*$, the conditional law of $\mathbf{s}$ given t is the uniform distribution on the unit sphere of $(\mathbf{c}^*)^{\perp}$, which is invariant under every rotation fixing $\mathbf{c}^*$; the only fixed point of all such rotations is the origin, so

$$\mathbb{E}[\mathbf{s}|t] = \mathbf{0}, \qquad \text{hence} \qquad \mathbb{E}[\mathbf{s}] = \mathbf{0}. \tag{20}$$

Suppose the weight depends on $\mathbf{u}$ only through t, that is there exists a function $\tilde{w} : [-1,1] \to \mathbb{R}_{\geq 0}$ with $w(\mathbf{u}) = \tilde{w}(t)$. This holds for cosine with $\tilde{w}(t) = \rho t$ (from the limit above), and for the calibrated probability model with $\tilde{w} = g$, where g is the increasing likelihood profile introduced below. Substituting Eq. 19 and using Eq. 20,

$$\mathbb{E}[w(\mathbf{u})\mathbf{u}] = \underbrace{\mathbb{E}[\tilde{w}(t)\,t]}_{>0}\,\mathbf{c}^* + \mathbb{E}\big[\tilde{w}(t)\sqrt{1-t^2}\,\mathbb{E}[\mathbf{s}|t]\big] = \mathbb{E}[\tilde{w}(t)\,t]\,\mathbf{c}^*, \tag{21}$$

where $\mathbb{E}[\tilde{w}(t)\,t] > 0$ because $\tilde{w}(t) \geq 0$ is a weight whose mass lies at $t > 0$ (near $\mathbf{c}^*$, since $\rho > 0$). Therefore $\mathbf{m} = \big(\mathbb{E}[\tilde{w}(t)\,t]/\mathbb{E}[\tilde{w}(t)]\big)\mathbf{c}^*$ is a positive multiple of $\mathbf{c}^*$, and $\mathbf{c}(P)/\|\mathbf{c}(P)\|_2 \xrightarrow{a.s.} \mathbf{c}^*$.

On the probability weight. The step $w(\mathbf{u}) = \tilde{w}(t)$ for the probability variant is a *calibration assumption*, not a fact: we assume the likelihood factors through alignment to the semantic center, $\Pr(y|x) = g(t)$ for some increasing $g : [-1,1] \to \mathbb{R}_{\geq 0}$, so that answers closer to $\mathbf{c}^*$ (larger t) receive larger weight. When calibration fails, $\mathbf{m}$ need not point along $\mathbf{c}^*$; we therefore treat the probability variant's guarantee as conditional on this assumption.

Remark (necessity of symmetry). Eq. 20 is the only place rotational symmetry is used: it forces the orthogonal component of $\mathbb{E}[w(\mathbf{u})\mathbf{u}]$ to vanish, leaving $\mathbf{m}$ parallel to $\mathbf{c}^*$. If $\mathcal{D}$ is skewed off axis, then $\mathbb{E}[\mathbf{s}] \neq \mathbf{0}$, the orthogonal term no longer cancels, and $\mathbf{m}$ tilts away from $\mathbf{c}^*$, so the cosine and probability variants lose their directional guarantee. The geometric group is unaffected, as its weights never depend on $\mathbf{u}$.

Conclusion: probability of correct selection. Since $\mathbf{c}(P)/\|\mathbf{c}(P)\|_2 \xrightarrow{a.s.} \mathbf{c}^*$ in all four variants, the selected candidate satisfies

$$\hat{i} = \arg\max_i \langle \mathbf{u}_i, \mathbf{c}(P)\rangle \xrightarrow{a.s.} \arg\max_i \langle \mathbf{u}_i, \mathbf{c}^*\rangle, \tag{22}$$

i.e. RCS selects the candidate best aligned with the semantic center, and

$$\Pr(\text{RCS selects the semantically correct answer}) \to 1 \qquad \text{as } N \to \infty. \tag{23}$$

Exact-match voting, by contrast, operates on surface strings. When the correct semantic cluster a^* is split across $L \geq 2$ surface forms (e.g. "Canberra", "It's Canberra", "Canberra ACT") with no single form holding a plurality, a compact wrong cluster with one surface form can win. With $L \geq 2$ fragments of a^*, no fragment attains plurality and voting selects a^* with probability at most $1/L$, bounded away from 1. Thus, as $N \to \infty$, RCS attains probability 1 while voting remains capped at $1/L$: RCS is asymptotically superior whenever surface fragmentation occurs. □

A.3 Proof of Proposition 2

Setup. The N candidates split into cluster A (correct), of size k, with mean direction $\boldsymbol{\mu}_A \in \mathbb{S}^{d-1}$, and cluster B (spurious), of size $N-k$, with mean direction $\boldsymbol{\mu}_B \in \mathbb{S}^{d-1}$; each candidate satisfies $\mathbb{E}[\mathbf{u}_i] = \rho\boldsymbol{\mu}_A$ for $i \in A$ and $\mathbb{E}[\mathbf{u}_i] = \rho\boldsymbol{\mu}_B$ for $i \in B$, with the same concentration $\rho \in (0, 1]$ as in Assumption 9. Let $\theta = \angle(\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)$.

Step 1: the bisector direction. RCS assigns a candidate $\mathbf{u}$ to the cluster whose mean direction it aligns with more, that is to A iff $\langle \mathbf{u}, \boldsymbol{\mu}_A\rangle > \langle \mathbf{u}, \boldsymbol{\mu}_B\rangle$, equivalently $\langle \mathbf{u}, \boldsymbol{\mu}_B - \boldsymbol{\mu}_A\rangle < 0$. Define the unit *bisector direction*

$$\mathbf{v} := \frac{\boldsymbol{\mu}_B - \boldsymbol{\mu}_A}{\|\boldsymbol{\mu}_B - \boldsymbol{\mu}_A\|_2}, \qquad \|\boldsymbol{\mu}_B - \boldsymbol{\mu}_A\|_2 = \sqrt{2 - 2\cos\theta} = 2\sin(\theta/2). \tag{24}$$

The vector $\mathbf{v}$ points from the correct cluster toward the spurious one, and the entire cluster decision is governed by the single scalar $\langle \mathbf{c}(P), \mathbf{v}\rangle$, with negative values favoring A. This is a projection onto one fixed direction: the value of $\langle \mathbf{c}(P), \mathbf{v}\rangle$ does not depend on the ambient dimension d, so no dimension-dependent variance enters the decision (in contrast to comparing the full d-dimensional distances, whose orthogonal components would contribute noise growing with d).

Step 2: expected margin. For the geometric group with uniform weights, $\mathbf{c}(P) = \frac{1}{N}\sum_i \mathbf{u}_i$ and

$$\mathbb{E}[\mathbf{c}(P)] = \frac{k}{N}\,\rho\boldsymbol{\mu}_A + \frac{N-k}{N}\,\rho\boldsymbol{\mu}_B. \tag{25}$$

Each mean direction lies at angle $\theta/2$ from the bisector, on opposite sides, so $\langle \boldsymbol{\mu}_A, \mathbf{v}\rangle = -\sin(\theta/2)$ and $\langle \boldsymbol{\mu}_B, \mathbf{v}\rangle = +\sin(\theta/2)$. Hence

$$\langle \mathbb{E}[\mathbf{c}(P)], \mathbf{v}\rangle = \rho\sin(\tfrac{\theta}{2})\Big(\tfrac{N-k}{N} - \tfrac{k}{N}\Big) = -\rho\sin(\tfrac{\theta}{2})\,\frac{2k-N}{N} = -\Delta, \qquad \Delta := \rho\sin(\tfrac{\theta}{2})\frac{2k-N}{N}. \tag{26}$$

The expected center projects to $-\Delta$ along $\mathbf{v}$, lying on the correct side of the bisector exactly when $\Delta > 0$, that is (uniform) when $k > N/2$. This restates the condition $k > N/2$ as a signed margin.

Step 3: concentration. Write the projected center as an average of scalars,

$$\langle \mathbf{c}(P), \mathbf{v}\rangle = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad X_i := \langle \mathbf{u}_i, \mathbf{v}\rangle. \tag{27}$$

The X_i are independent (the candidates are drawn independently), and each is bounded, $X_i \in [-1, 1]$, because $\|\mathbf{u}_i\|_2 = \|\mathbf{v}\|_2 = 1$ and the inner product of two unit vectors lies in $[-1, 1]$. Their average has mean $\mathbb{E}[\frac{1}{N}\sum_i X_i] = \langle \mathbb{E}[\mathbf{c}(P)], \mathbf{v}\rangle = -\Delta$. RCS selects A when $\frac{1}{N}\sum_i X_i < 0$, i.e. when the average exceeds its mean $-\Delta$ by less than Δ. Hoeffding's inequality (Hoeffding, 1963), for independent variables each bounded in an interval of length $b_i - a_i = 2$, gives for any $t > 0$

$$\Pr\Big(\tfrac{1}{N}\textstyle\sum_i X_i - \mathbb{E}\big[\tfrac{1}{N}\sum_i X_i\big] \geq t\Big) \leq \exp\Big(-\frac{2N^2t^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\Big) \tag{28}$$

$$= \exp\Big(-\frac{2N^2t^2}{4N}\Big) = \exp\Big(-\frac{Nt^2}{2}\Big). \tag{29}$$

Setting $t = \Delta > 0$,

$$\Pr(\text{RCS selects } A) = \Pr\Big(\tfrac{1}{N}\textstyle\sum_i X_i < 0\Big) \tag{30}$$

$$= 1 - \Pr\Big(\tfrac{1}{N}\textstyle\sum_i X_i - (-\Delta) \geq \Delta\Big) \tag{31}$$

$$\geq 1 - \exp\big(-N\Delta^2/2\big). \tag{32}$$

The bound is non-asymptotic, depends only on N and Δ and not on d, and requires no distributional assumption beyond $\Delta > 0$.

Step 4: other weighting variants. *Frequency.* Replacing uniform weights by $p_i = f_i / \sum_j f_j$ changes the deterministic part of the projected center to $\rho \sin(\frac{\theta}{2})(W_B - W_A)$, where W_A, W_B are the total normalized frequency mass of clusters A and B. The same Hoeffding step applies with margin $\Delta_{\text{freq}} = \rho \sin(\frac{\theta}{2})(W_A - W_B)$, so a lexically dominant correct cluster ($W_A > W_B$) is recovered even when $k < N/2$.

Cosine. Endogenous weights amplify whichever cluster is most concentrated; when a spurious cluster is tighter than the correct one, the projected margin can turn negative even at $k > N/2$, so no finite sample guarantee holds.

Probability. Exogenous weights add a term proportional to the likelihood gap between the clusters. When the likelihood favors the correct cluster, this term is positive and adds to Δ, so the correct cluster can be selected even at $k < N/2$, the one case in which RCS recovers a genuine semantic minority. As this depends on external likelihood calibration, we state it as a mechanism rather than a guarantee.

Remark. When $\Delta < 0$ (a semantically isolated correct minority under the geometric group), the expected center lies on cluster B's side and RCS selects the wrong cluster, failing exactly as majority voting does. Asymptotic recovery (Proposition 1) is preserved for all k as $N \to \infty$. □

A.4 Limitations

While efficient, RCS relies on embedding quality and assumes that correct answers cluster around a semantic centroid, which may cause semantically isolated correct answers to be overlooked when supported by only a small minority. This assumption may be less effective with diverse valid answers or tightly clustered hallucinations from shared reasoning flaws. Improving embedding robustness and better handling such diversity are promising directions for future work.

A.5 Implementation Details

Evaluation Protocol. For all methods, we measure accuracy (%) between the selected answer and the ground-truth. A generation is deemed correct via exact match on long-form reasoning tasks or ROUGE-L F1 > 0.3 on short-form QA tasks (SciQ, GPQA), exactly as in prior work (Kuhn et al., 2023; Nguyen et al., 2026a). We include all model outputs, including empty or degenerate responses, to reflect real-world noisy generation settings.

We use 5-shot prompting (Brown et al., 2020) for short-form QA and Chain-of-Thought prompting (Wei et al., 2022) for long-form tasks. For all baselines, we follow the original setup. We summarize the evaluation benchmarks, including the number of evaluation samples and representative examples, in Table 7. For MMLU-Pro, we sample up to 10 questions per category (105 samples across 14 categories; some categories contain fewer than 10 eligible questions) to ensure broad coverage; for BigBenchHard, we consider two challenging tasks: Date Understanding (**BBH-Date**) and Navigate (**BBH-Navigate**). We observe no differences between breaking ties randomly and preserving the original answer order. For determinism and reproducibility, we adopt the latter. For the multi-agent debate setting, we follow prior implementations (Choi et al., 2025; Nguyen et al., 2026b). Sampling hyperparameters and generation prompts are provided in Tables 5 and 6, respectively. All experiments are conducted with three different seeds (42, 44, and 46) on a single H100-80GB GPU. Total compute across all models and benchmarks is approximately 120 GPU-hours; a single model evaluation (one benchmark, one model, N=10) completes in roughly 2–5 hours depending on model size.

Table 5: Sampling hyperparameters used for generation.

Parameter	Value
Temperature	1
Max new tokens (short-form QA)	32
Max new tokens (long-form tasks)	512

Table 6: System prompts used for generation across tasks.

Tasks	Prompt
SciQ, GPQA	This is a bot that correctly answers questions.
Arithmetics, GSM8K	Make sure to state your final answer in curly brackets at the very end of your response, just like: `{final answer: 12.34}`. Let's think step by step.
Form.Log., MMLU-Pro, BBH-Date	Make sure to state your final answer choice in curly brackets at the very end of your response, just like: `{final answer: (A)}`. Let's think step by step.
BBH-Navigate	Make sure to state your final answer choice in curly brackets at the very end of your response, just like: `{final answer: Yes}`. Let's think step by step.

Table 7: Overview of evaluation benchmarks, including the number of samples and representative examples.

Dataset	#Samples	Question	Answer
SciQ	1000	Compounds that are capable of accepting electrons, such as o 2 or f2, are called what?	oxidants
GPQA	198	Which of the following physical theories never requires UV regularization?	Superstring Theory
Arithmetics	200	What is $23 + 47$?	70
GSM8K	200	A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?	3
Formal Logic	126	Select the best translation into predicate logic: Sheena is a punk rocker. *Choices:* ("Sx", "xS", "sP", "Ps")	"Ps"
MMLU-Pro	105	Determine the number of men needed to build a boat in 77 days if it takes 36 men 132 days to build one. *Choices:* ("84 men", "36 men", "99 men", "132 men", "45 men", "70 men", "62 men", "50 men", "77 men", "120 men")	"62 men"
BBH-Date	250	Today is Christmas Eve of 1937. What is the date tomorrow in MM/DD/YYYY? *Options:* (A) 12/11/1937 (B) 12/25/1937 (C) 01/04/1938 (D) 12/04/1937 (E) 12/25/2006 (F) 07/25/1937	(B)
BBH-Navigate	250	If you follow these instructions, do you return to the starting point? Always face forward. Take 1 step backward. Take 9 steps left. Take 2 steps backward. Take 6 steps forward. Take 4 steps forward. Take 4 steps backward. Take 3 steps right. *Options:* - Yes - No	No

A.6 Extended Results: Number Of Samples

Tables 10, 11, and 12 report performance over different number of sampling responses.

Table 8: Extended Results: Best-of-N selection accuracy across settings. Best-of-N selection accuracy. **BBF**: black-box friendly. For non-RCS methods, underline marks the best baseline per setting. Among RCS variants, **bold** indicates the best RCS method per column. No baseline significantly outperforms RCS at the column level ($p < 0.05$, paired t-test); a full per-cell (model $\times$ benchmark) McNemar breakdown against all six baselines is in Appendix Table 9.

Model	Method	BBF	SciQ	GPQA	Arithmetics	GSM8K	Form.Log.	Average
Qwen2.5-7B	Greedy	✓	64.1	20.5	97.0	88.3	43.6	63.1
	Oracle	✓	82.0±1.1	49.9±1.6	100.0±0.0	94.6±0.8	67.7±2.5	78.8
	NLL	✗	66.4±0.1	18.3±1.3	47.3±5.7	32.7±4.3	25.7±0.5	38.1
	ANLL	✗	66.2±0.4	18.1±0.9	47.0±5.6	32.7±4.4	25.9±0.5	38.0
	SC	✓	70.3±0.9	22.1±4.2	74.3±2.0	52.3±2.0	48.4±0.0	53.5
	SCW	✓	70.2±0.8	26.3±3.2	52.8±5.0	36.4±5.2	32.8±3.3	43.7
	CE	✗	70.4±0.4	21.6±3.8	75.5±4.0	51.9±3.1	48.7±1.2	53.6
	ModeX	✓	69.0±0.6	21.6±1.6	54.3±5.1	37.6±3.3	35.2±2.6	43.5
	RCS_{medoid}	✓	70.1±0.6	25.0±1.3	77.5±3.6	57.5±2.6	**49.2±2.9**	55.9
	RCS_{uni}	✓	70.2±0.2	**27.1±1.7**	77.5±2.3	61.6±2.8	**49.2±3.2**	57.1
	RCS_{freq}	✓	**70.3±0.5**	25.9±1.8	76.7±1.4	56.0±2.1	47.9±2.4	55.4
	RCS_{cosine}	✓	70.1±0.5	26.9±1.8	76.8±3.4	61.3±3.1	**49.2±3.2**	56.9
	RCS_{prob}	✗	70.0±0.4	**27.1±2.4**	**77.7±1.9**	**62.1±3.1**	**49.2±3.2**	**57.2**
Llama3.1-8B	Greedy	✓	63.1	25.5	89.5	88.3	34.9	60.8
	Oracle	✓	85.1±0.4	56.3±1.3	99.3±0.3	94.6±0.8	74.1±1.2	81.9
	NLL	✗	61.0±0.5	20.4±0.8	73.3±2.8	69.0±0.5	33.9±2.8	51.5
	ANLL	✗	61.1±0.5	18.7±0.9	73.5±4.1	68.7±1.6	33.1±2.4	51.0
	SC	✓	67.3±0.7	16.8±1.8	78.3±3.0	66.0±2.3	31.7±2.1	52.0
	SCW	✓	67.7±1.0	26.8±0.8	53.8±4.2	50.3±5.9	25.7±1.8	44.9
	CE	✗	66.8±0.9	18.5±1.1	81.0±3.8	67.5±1.3	32.3±1.7	53.2
	ModeX	✓	63.1±1.2	17.8±1.1	55.3±2.5	48.4±3.9	25.1±2.6	41.9
	RCS_{medoid}	✓	68.0±0.9	26.4±0.9	86.5±1.5	71.4±2.8	37.8±0.9	58.0
	RCS_{uni}	✓	**68.1±0.7**	26.9±1.0	**88.5±0.5**	72.8±3.0	**39.4±0.9**	**59.1**
	RCS_{freq}	✓	67.6±1.1	26.9±1.0	83.7±2.3	68.9±1.9	34.9±0.8	56.4
	RCS_{cosine}	✓	**68.1±0.7**	27.3±0.8	87.7±1.2	72.8±2.3	**39.4±0.9**	**59.1**
	RCS_{prob}	✗	67.9±1.1	**27.5±0.9**	85.2±0.6	**75.6±1.0**	36.8±2.6	58.6
Gemma2-9B	Greedy	✓	67.6	22.6	96.0	69.0	52.4	62.0
	Oracle	✓	84.6±0.6	48.5±4.2	98.2±0.6	95.3±0.3	77.8±5.7	80.9
	NLL	✗	72.9±1.1	25.6±1.1	94.8±1.3	67.9±1.1	48.1±2.8	61.9
	ANLL	✗	66.3±0.9	21.2±2.3	96.0±1.0	73.1±0.9	50.0±0.8	61.3
	SC	✓	73.9±0.2	19.7±0.9	97.3±0.3	72.9±3.5	51.6±0.8	63.1
	SCW	✓	74.0±0.6	25.4±0.5	92.2±1.0	60.1±1.8	48.9±3.2	60.1
	CE	✗	73.6±0.6	18.6±0.5	97.5±0.0	73.1±3.7	52.1±1.7	63.0
	ModeX	✓	71.7±0.2	19.5±2.2	92.7±2.8	62.6±1.5	49.2±4.2	59.1
	RCS_{medoid}	✓	**74.2±0.4**	24.4±1.4	97.2±0.6	72.9±3.5	52.6±0.9	64.3
	RCS_{uni}	✓	74.0±0.4	25.9±0.5	97.2±0.6	73.8±2.8	**52.9±2.0**	64.8
	RCS_{freq}	✓	73.9±0.5	25.7±0.6	**97.3±0.3**	72.6±3.7	51.8±1.2	64.3
	RCS_{cosine}	✓	74.1±0.4	25.7±0.3	97.2±0.6	71.1±4.5	51.8±1.2	64.0
	RCS_{prob}	✗	**74.2±0.5**	**26.3±0.6**	**97.3±0.3**	**75.1±3.5**	52.4±0.8	**65.1**

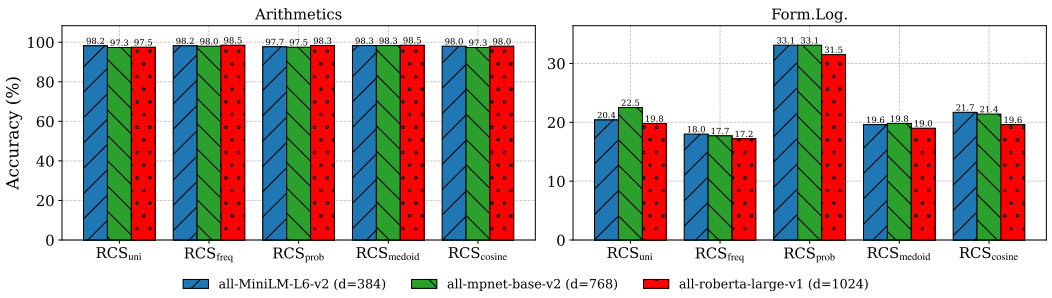

Figure 6: Effect of encoders on Llama3.2-3B. Results for Qwen2.5-3B are in Appendix Figure 7.

Table 9: Full per-cell (model × benchmark) item-level McNemar p-value of RCS_{uni} against each of the six baselines. **Bold** = significant at α=0.05 in RCS_{uni}'s favor; **red bold** = significant in the baseline's favor.

Model	Baseline	SciQ	GPQA	Arithmetics	GSM8K	Form.Log.
Qwen2.5-3B	NLL	$< 10^{-20}$	$< 10^{-4}$	$< 10^{-74}$	$< 10^{-54}$	$< 10^{-16}$
	ANLL	$< 10^{-21}$	$< 10^{-4}$	$< 10^{-74}$	$< 10^{-54}$	$< 10^{-17}$
	SC	**0.001**	0.326	$< 10^{-7}$	$< 10^{-18}$	**0.003**
	SCW	**0.001**	1.000	0.108	$< 10^{-6}$	0.388
	CE	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-19}$	**0.036**
	ModeX	$< 10^{-9}$	**0.007**	$< 10^{-39}$	$< 10^{-28}$	$< 10^{-6}$
Qwen2.5-7B	NLL	$< 10^{-11}$	$< 10^{-6}$	$< 10^{-29}$	$< 10^{-27}$	$< 10^{-18}$
	ANLL	$< 10^{-13}$	$< 10^{-6}$	$< 10^{-29}$	$< 10^{-27}$	$< 10^{-18}$
	SC	0.901	**0.002**	$< 10^{-4}$	$< 10^{-13}$	0.815
	SCW	0.766	0.839	0.481	$< 10^{-4}$	1.000
	CE	0.665	**0.001**	**0.029**	$< 10^{-11}$	1.000
	ModeX	**0.011**	**0.003**	$< 10^{-15}$	$< 10^{-17}$	$< 10^{-10}$
Llama3.2-3B	NLL	$< 10^{-16}$	**0.009**	$< 10^{-16}$	**0.000**	**$< 10^{-6}$**
	ANLL	$< 10^{-16}$	**0.004**	$< 10^{-12}$	$< 10^{-4}$	**0.000**
	SC	0.080	$< 10^{-5}$	0.344	**0.002**	**0.000**
	SCW	1.000	0.791	1.000	0.096	**0.002**
	CE	0.750	**0.022**	0.227	**0.011**	**0.012**
	ModeX	$< 10^{-16}$	$< 10^{-4}$	$< 10^{-40}$	$< 10^{-19}$	0.091
Llama3.1-8B	NLL	$< 10^{-24}$	**0.000**	$< 10^{-12}$	0.078	**0.018**
	ANLL	$< 10^{-23}$	$< 10^{-5}$	$< 10^{-12}$	0.053	**0.008**
	SC	0.056	$< 10^{-7}$	$< 10^{-15}$	$< 10^{-10}$	$< 10^{-6}$
	SCW	0.201	1.000	**0.000**	**0.001**	**0.003**
	CE	**0.003**	$< 10^{-6}$	$< 10^{-8}$	$< 10^{-5}$	$< 10^{-5}$
	ModeX	$< 10^{-21}$	$< 10^{-4}$	$< 10^{-44}$	$< 10^{-21}$	$< 10^{-7}$
Gemma2-9B	NLL	**0.016**	0.902	**0.001**	**0.006**	**0.012**
	ANLL	$< 10^{-37}$	**0.003**	0.065	1.000	0.132
	SC	0.834	$< 10^{-4}$	1.000	0.500	0.125
	SCW	1.000	0.581	1.000	**0.007**	0.125
	CE	0.337	$< 10^{-5}$	0.500	1.000	0.625
	ModeX	$< 10^{-6}$	$< 10^{-4}$	$< 10^{-6}$	$< 10^{-5}$	0.542

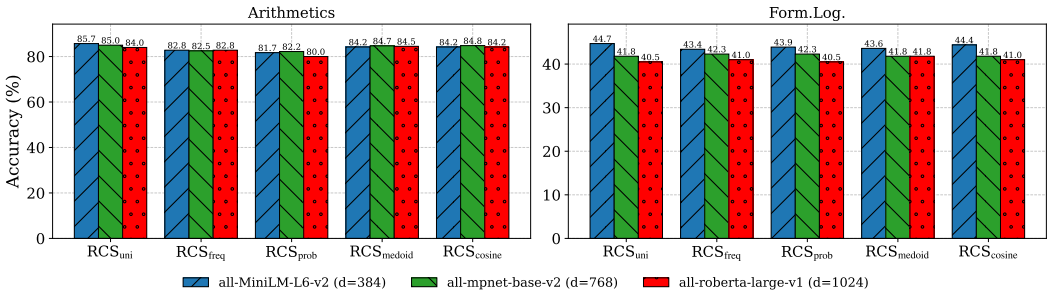

Figure 7: Effect of the sentence embedding model on Arithmetics and Form.Log. using Qwen2.5-3B.

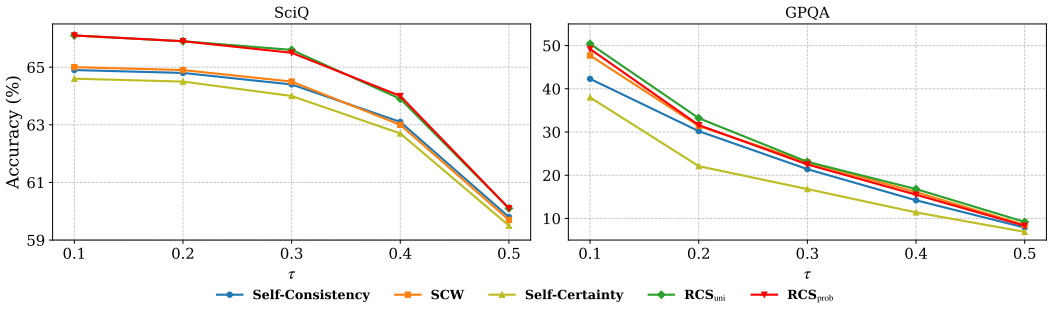

Figure 8: Performance when varying correctness threshold (τ) using Llama3.2-3B.

Table 10: Best-of-N selection accuracy across settings (N=5).

Model	Method	SciQ	GPQA	Arithmetics	GSM8K	Form.Log.
Qwen2.5-3B	NLL	61.7±1.2	17.6±3.1	49.2±12.4	45.2±4.7	28.0±3.2
	ANLL	61.6±1.1	17.1±4.1	48.7±11.4	46.0±4.8	27.5±3.3
	SC	63.8±0.5	18.1±2.7	78.7±7.2	65.0±0.5	39.4±1.8
	SCW	64.0±0.2	24.0±2.6	46.8±1.1	38.1±6.2	27.0±2.1
	CE	63.6±0.3	15.7±1.7	78.5±7.4	66.8±5.0	42.4±0.2
	ModeX	61.2±1.2	18.0±2.4	46.0±3.0	32.0±0.5	24.9±0.9
	RCS_{medoid}	64.0±0.4	24.7±2.2	77.7±2.8	66.7±2.9	40.7±0.5
	RCS_{uni}	63.8±0.1	24.2±2.9	77.0±2.8	66.5±4.0	40.7±0.5
	RCS_{freq}	64.0±0.2	24.0±2.6	77.5±3.1	66.3±3.8	41.0±0.5
	RCS_{cosine}	63.9±0.2	24.9±2.4	72.8±2.5	55.5±2.4	36.5±1.4
	RCS_{prob}	64.0±0.5	21.6±2.0	75.7±5.3	63.1±4.4	41.3±0.0
Qwen2.5-7B	NLL	67.6±1.5	17.1±1.4	63.3±10.9	51.4±3.4	36.0±2.3
	ANLL	67.5±1.1	16.9±1.2	63.5±11.0	51.4±3.4	35.4±2.0
	SC	69.2±0.5	22.3±3.2	83.3±13.7	72.4±3.1	47.4±4.4
	SCW	68.9±0.4	26.9±0.9	53.2±1.1	38.9±2.8	32.0±1.2
	CE	69.4±0.1	19.9±1.1	84.0±16.0	74.3±2.6	47.8±3.6
	ModeX	69.0±0.7	22.1±3.0	54.2±3.2	40.6±2.7	34.1±6.3
	RCS_{medoid}	69.0±0.4	27.3±0.8	84.3±12.0	74.3±1.8	47.6±3.6
	RCS_{uni}	68.9±0.4	27.6±1.3	84.2±11.8	75.0±1.3	47.6±3.6
	RCS_{freq}	69.0±0.4	26.9±0.9	84.3±12.0	75.1±1.0	47.6±3.6
	RCS_{cosine}	69.0±0.4	27.5±1.0	74.8±6.0	60.1±2.4	45.2±4.4
	RCS_{prob}	69.3±0.4	25.2±1.5	83.0±11.8	70.9±2.0	47.9±3.3
Llama3.2-3B	NLL	54.0±1.3	19.2±0.9	82.7±4.2	73.4±3.8	31.0±2.9
	ANLL	53.7±0.6	18.5±2.0	86.7±3.3	76.1±1.8	31.5±2.0
	SC	58.6±0.8	16.1±2.4	93.3±2.1	79.2±3.6	31.5±1.7
	SCW	59.4±1.2	22.3±2.6	69.8±0.4	59.7±1.3	16.9±3.9
	CE	57.6±1.4	19.2±3.2	93.0±1.8	80.3±4.1	33.7±2.4
	ModeX	53.4±1.1	12.8±3.2	68.8±1.0	61.3±2.6	19.1±3.5
	RCS_{medoid}	59.4±1.2	22.5±2.9	94.0±0.5	80.2±2.0	30.9±4.8
	RCS_{uni}	59.3±1.5	22.3±2.6	93.3±0.6	80.4±2.1	30.9±4.8
	RCS_{freq}	59.4±1.2	22.3±2.6	93.8±0.8	80.4±2.1	30.9±4.8
	RCS_{cosine}	59.3±1.4	22.5±2.9	92.8±1.1	76.3±1.6	20.9±4.4
	RCS_{prob}	58.9±1.4	21.2±3.6	93.5±1.5	80.0±1.9	32.8±2.4
Llama3.1-8B	NLL	61.2±0.6	21.1±1.1	73.3±2.5	72.8±3.3	38.9±1.4
	ANLL	60.8±0.5	20.7±1.4	73.2±3.3	73.3±3.1	38.4±1.8
	SC	66.0±1.8	19.2±1.9	82.2±7.2	82.6±3.8	43.1±3.9
	SCW	67.1±0.7	26.8±1.3	52.8±3.3	55.7±3.1	28.6±3.6
	CE	65.6±0.4	19.7±3.2	82.1±8.4	82.4±3.1	42.4±3.0
	ModeX	62.1±0.9	20.5±2.1	50.2±0.8	53.5±3.8	25.1±0.5
	RCS_{medoid}	67.0±0.8	27.3±1.3	86.0±4.0	83.2±5.2	41.8±3.2
	RCS_{uni}	67.0±0.6	26.9±2.1	85.5±3.0	83.2±5.6	42.3±3.6
	RCS_{freq}	67.1±0.7	26.8±1.3	85.7±3.6	83.2±5.6	42.3±3.6
	RCS_{cosine}	66.9±0.6	27.1±2.1	81.3±0.8	70.9±2.4	36.8±2.6
	RCS_{prob}	66.7±0.8	26.4±2.7	77.3±3.6	76.0±3.1	39.2±4.1
Gemma2-9B	NLL	69.2±0.4	27.1±1.1	94.5±1.0	87.3±0.9	54.5±4.4
	ANLL	66.2±0.9	21.1±1.1	95.8±1.0	87.8±1.3	54.8±2.1
	SC	72.8±0.6	17.3±2.3	97.0±0.9	88.8±1.8	55.6±5.0
	SCW	73.9±0.6	24.9±2.3	92.0±1.8	65.7±2.5	47.6±5.6
	CE	73.2±0.4	17.8±3.5	97.0±0.9	89.8±2.1	55.5±5.2
	ModeX	71.5±0.6	20.7±0.5	92.5±1.0	64.0±2.3	52.1±6.5
	RCS_{medoid}	73.7±0.7	24.2±2.0	96.8±0.6	89.0±2.1	55.3±4.8
	RCS_{uni}	73.8±0.4	25.9±2.1	96.8±0.6	88.7±1.6	55.8±4.8
	RCS_{freq}	74.0±0.6	24.9±2.3	96.8±0.6	88.7±1.6	55.6±4.4
	RCS_{cosine}	73.8±0.5	25.7±2.2	96.8±0.6	72.8±1.1	54.0±4.2
	RCS_{prob}	73.2±0.3	24.2±1.3	96.8±0.6	88.8±1.5	55.3±3.6

Table 11: Best-of-N selection accuracy across settings (N=20).

Model	Method	SciQ	GPQA	Arithmetics	GSM8K	Form.Log.
Qwen2.5-3B	NLL	59.0±0.4	16.4±2.3	48.2±11.9	40.6±0.5	26.5±2.6
	ANLL	58.4±0.4	16.2±2.6	47.8±12.5	40.8±1.1	26.5±2.6
	SC	64.9±0.6	20.0±0.3	95.3±5.1	88.3±0.9	48.1±3.6
	SCW	64.9±0.6	24.4±2.7	56.2±4.6	33.8±3.9	28.3±1.8
	CE	64.2±0.9	18.3±1.3	95.3±5.9	88.8±0.9	47.1±2.4
	ModeX	62.2±0.8	21.9±1.7	51.7±3.4	37.9±2.1	23.3±1.8
	RCS_{medoid}	65.0±0.7	26.1±2.7	96.5±3.5	88.5±0.6	47.4±4.1
	RCS_{uni}	65.5±1.0	24.9±2.7	96.7±2.4	87.1±1.6	48.1±4.4
	RCS_{freq}	64.7±0.5	23.8±2.9	96.3±3.3	88.0±0.8	47.6±5.0
	RCS_{cosine}	65.3±0.9	25.2±2.9	94.2±1.1	70.4±1.8	42.3±3.9
	RCS_{prob}	65.3±0.7	24.9±2.4	95.2±3.3	86.6±1.1	46.6±4.1
Qwen2.5-7B	NLL	66.5±0.8	18.1±1.9	63.3±11.2	55.8±2.0	30.7±4.8
	ANLL	66.5±1.0	18.1±1.8	63.5±12.2	55.7±1.8	31.2±4.4
	SC	70.0±0.2	26.3±0.6	93.3±9.8	93.1±1.3	51.8±2.3
	SCW	70.0±0.3	27.6±2.0	56.2±1.1	37.2±2.6	34.9±4.0
	CE	69.9±0.2	25.2±0.3	93.7±9.7	92.6±0.8	52.9±2.6
	ModeX	68.5±1.2	24.0±2.6	59.5±0.0	37.9±1.5	31.2±3.6
	RCS_{medoid}	69.7±0.3	28.0±1.8	95.0±6.9	92.2±0.8	52.6±1.7
	RCS_{uni}	69.9±0.2	28.7±2.2	95.8±5.9	91.7±1.1	53.4±2.4
	RCS_{freq}	70.1±0.3	28.0±1.6	94.5±7.8	92.7±1.1	52.1±1.8
	RCS_{cosine}	69.8±0.2	28.3±1.8	86.0±2.8	70.4±2.4	48.7±1.7
	RCS_{prob}	70.0±0.3	27.5±1.6	95.3±6.8	91.7±0.3	53.4±3.2
Llama3.2-3B	NLL	53.0±0.5	22.1±2.9	90.2±0.8	79.9±2.1	39.7±0.0
	ANLL	52.8±0.9	19.3±3.5	91.7±0.3	82.7±0.9	39.9±5.3
	SC	59.8±0.3	15.7±0.8	98.2±0.3	89.3±1.3	43.7±2.9
	SCW	60.3±0.4	25.0±2.6	67.8±1.8	58.2±6.1	20.9±2.0
	CE	60.0±0.6	17.3±1.5	98.2±0.3	89.7±1.3	45.0±3.0
	ModeX	55.0±0.7	16.9±3.2	73.0±0.5	61.8±5.7	20.1±2.0
	RCS_{medoid}	60.7±0.2	26.1±3.4	98.0±0.0	89.5±1.3	43.9±3.2
	RCS_{uni}	60.7±0.7	26.8±2.9	98.5±0.5	89.5±1.6	44.2±2.4
	RCS_{freq}	60.4±0.3	25.7±2.9	98.0±0.0	89.3±1.3	43.4±2.5
	RCS_{cosine}	60.8±0.6	26.1±4.0	98.5±0.0	84.3±1.5	22.8±1.7
	RCS_{prob}	60.7±0.5	26.4±3.7	98.0±0.0	89.5±1.1	42.1±2.4
Llama3.1-8B	NLL	61.0±1.5	20.7±1.5	81.2±4.3	83.8±1.8	48.9±1.8
	ANLL	61.1±1.9	21.8±0.0	79.8±3.9	82.2±3.1	44.4±2.1
	SC	68.1±0.5	18.9±0.4	95.3±6.0	91.9±1.3	55.3±2.0
	SCW	68.3±0.1	32.4±0.4	49.5±2.3	51.1±1.5	27.0±1.6
	CE	68.2±0.6	20.7±0.0	94.8±7.2	91.7±1.5	54.0±2.1
	ModeX	62.8±0.7	22.3±0.5	54.8±0.6	49.9±3.6	24.3±3.3
	RCS_{medoid}	67.9±0.1	35.8±0.7	97.8±2.1	92.2±1.8	55.0±2.4
	RCS_{uni}	68.3±0.5	36.0±1.1	98.8±1.2	91.4±1.8	54.0±2.4
	RCS_{freq}	68.3±0.5	33.7±0.7	96.0±5.2	92.2±1.8	55.0±2.4
	RCS_{cosine}	68.3±0.3	36.0±2.6	95.3±1.4	83.6±1.1	39.4±3.7
	RCS_{prob}	68.1±0.3	36.3±0.7	97.2±1.2	88.3±1.8	52.9±3.0
Gemma2-9B	NLL	74.1±0.3	27.1±1.1	96.2±0.3	89.3±0.9	52.1±0.9
	ANLL	64.8±0.5	23.0±1.1	96.3±0.3	89.8±0.9	54.5±1.7
	SC	74.1±0.8	22.1±2.0	97.2±0.3	91.7±0.3	54.0±0.8
	SCW	74.1±0.5	28.8±2.0	93.2±0.6	63.6±2.0	50.5±1.7
	CE	74.3±0.4	21.8±0.5	97.2±0.3	91.7±0.8	54.0±0.8
	ModeX	73.0±0.4	22.1±1.6	92.3±2.0	64.8±1.1	51.9±5.4
	RCS_{medoid}	74.2±0.4	28.7±1.7	97.2±0.3	91.7±0.3	54.2±0.5
	RCS_{uni}	74.7±0.4	29.9±1.3	97.2±0.3	91.7±0.3	54.2±0.5
	RCS_{freq}	74.2±0.6	29.4±2.1	97.2±0.3	91.9±0.0	54.2±0.5
	RCS_{cosine}	74.6±0.4	29.0±0.5	97.2±0.3	76.3±0.8	52.1±0.9
	RCS_{prob}	74.6±0.6	29.5±0.9	97.2±0.3	91.5±0.3	54.2±0.9

Table 12: Best-of-N selection accuracy across settings (N=40).

Model	Method	SciQ	GPQA	Arithmetics	GSM8K	Form.Log.
Qwen2.5-3B	NLL	56.3±0.8	16.8±1.8	45.3±13.0	42.8±1.5	25.9±3.8
	ANLL	56.1±0.9	16.8±2.0	44.8±13.3	43.0±2.8	25.1±3.8
	SC	65.2±0.4	21.8±1.4	97.3±2.9	90.2±1.1	49.5±1.2
	SCW	65.3±0.3	25.6±2.3	54.5±4.2	38.1±1.8	30.7±3.0
	CE	65.2±0.3	20.4±2.7	97.5±2.6	90.4±0.5	48.9±3.2
	ModeX	63.0±1.0	20.0±0.5	49.8±3.2	35.7±3.1	25.7±6.7
	RCS_{medoid}	65.5±0.5	25.0±1.3	98.2±1.5	90.0±0.8	47.9±1.2
	RCS_{uni}	65.7±0.7	25.0±1.3	98.5±1.3	89.7±1.3	48.4±0.8
	RCS_{freq}	65.3±0.5	23.8±2.4	97.7±2.4	90.0±0.8	47.6±2.1
	RCS_{cosine}	65.7±0.7	25.7±0.6	97.2±1.1	76.1±2.2	44.7±1.2
	RCS_{prob}	65.9±0.7	24.4±1.0	98.2±1.6	89.3±1.3	47.4±0.5
Qwen2.5-7B	NLL	66.0±0.4	18.5±1.2	63.0±10.1	57.5±3.1	30.7±4.4
	ANLL	65.9±0.6	18.5±0.8	62.8±9.3	57.5±3.4	30.4±3.2
	SC	70.5±0.3	26.1±1.6	96.2±5.8	94.4±0.5	54.0±0.8
	SCW	70.4±0.1	28.2±0.6	50.2±1.8	39.1±4.8	32.8±2.3
	CE	70.4±0.6	26.9±3.1	96.3±5.5	94.4±0.0	53.7±1.2
	ModeX	69.6±0.3	25.0±1.7	53.2±1.8	38.6±1.8	31.2±2.6
	RCS_{medoid}	70.0±0.1	28.0±0.9	97.2±4.0	94.8±0.3	53.2±2.1
	RCS_{uni}	70.3±0.4	29.0±0.9	98.2±2.3	94.6±0.3	52.6±3.7
	RCS_{freq}	70.5±0.2	27.5±0.5	96.5±5.2	94.6±0.3	52.9±3.2
	RCS_{cosine}	70.2±0.4	28.3±0.3	94.0±0.0	75.6±1.8	50.3±1.7
	RCS_{prob}	70.4±0.4	27.3±0.8	97.5±3.5	94.6±0.3	52.4±3.5
Llama3.2-3B	NLL	53.4±0.3	21.6±1.7	93.5±2.6	82.7±1.0	36.2±4.0
	ANLL	52.8±0.6	18.7±2.3	93.3±2.5	83.2±1.8	41.8±5.7
	SC	60.4±0.7	17.6±0.9	98.7±0.6	91.2±0.8	44.2±0.5
	SCW	60.6±0.9	24.7±2.4	69.0±4.2	58.0±2.6	19.6±3.0
	CE	60.2±0.4	21.2±3.2	98.8±0.3	91.2±0.6	43.1±0.9
	ModeX	55.8±0.9	19.0±2.9	72.5±2.6	57.0±4.1	22.0±3.2
	RCS_{medoid}	60.9±1.0	29.4±1.6	98.7±0.6	91.2±0.6	45.2±0.8
	RCS_{uni}	60.9±0.8	29.4±2.0	98.5±0.9	90.4±0.5	46.3±1.7
	RCS_{freq}	61.0±1.0	26.9±1.9	98.7±0.6	90.9±0.5	44.2±1.7
	RCS_{cosine}	61.1±1.1	29.9±2.0	98.8±1.1	86.3±0.5	23.0±2.9
	RCS_{prob}	61.0±0.6	29.4±2.1	98.3±0.3	90.7±0.3	44.7±2.8
Llama3.1-8B	NLL	61.8±0.9	18.9±1.1	83.7±1.2	83.9±2.5	46.8±2.1
	ANLL	61.3±0.8	20.5±1.8	84.0±1.7	81.7±2.0	43.4±4.0
	SC	69.3±0.3	19.7±0.7	96.3±4.2	93.2±0.8	51.9±2.0
	SCW	69.2±0.4	25.6±1.1	54.2±1.0	52.6±1.2	27.0±2.9
	CE	68.5±0.4	18.9±0.4	96.5±4.3	93.2±0.8	51.1±1.7
	ModeX	64.8±0.3	19.3±2.7	53.5±0.9	51.4±1.6	23.0±2.1
	RCS_{medoid}	69.1±0.3	31.6±0.7	98.5±0.5	92.4±0.5	51.1±1.2
	RCS_{uni}	69.6±0.2	31.4±0.4	98.7±1.0	92.2±0.3	50.5±1.2
	RCS_{freq}	69.2±0.8	28.2±1.8	96.7±3.6	93.1±0.6	51.1±1.2
	RCS_{cosine}	69.5±0.2	31.9±1.1	99.2±0.6	84.9±0.8	38.1±1.4
	RCS_{prob}	69.6±0.4	29.5±1.5	98.5±0.9	91.9±1.3	51.6±2.1
Gemma2-9B	NLL	74.2±0.1	26.8±2.1	96.0±0.5	89.7±0.8	51.3±3.9
	ANLL	62.2±1.5	20.7±1.6	96.2±0.3	89.5±1.8	53.7±2.0
	SC	74.2±0.4	22.5±0.8	97.5±0.0	91.7±0.8	55.6±0.8
	SCW	74.2±0.5	26.9±0.9	90.3±2.1	64.8±1.1	50.5±3.3
	CE	74.2±0.3	19.3±0.6	97.5±0.0	91.5±0.3	55.3±1.2
	ModeX	71.9±0.4	21.4±3.2	93.5±0.5	66.3±2.9	54.0±2.4
	RCS_{medoid}	74.4±0.4	29.4±1.2	97.5±0.0	91.7±0.8	55.0±0.9
	RCS_{uni}	74.4±0.8	29.7±1.6	97.5±0.0	91.5±0.6	54.8±0.8
	RCS_{freq}	74.2±0.4	28.0±1.0	97.5±0.0	91.7±0.8	55.3±1.2
	RCS_{cosine}	74.3±0.6	29.9±0.8	97.5±0.0	76.7±0.9	52.6±0.5
	RCS_{prob}	74.4±0.5	29.7±0.8	97.3±0.3	91.2±0.8	55.3±0.5