Tracing the Traces: Latent-Space Metrics for Efficient and Accurate Reasoning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Reasoning models rely on inference-time scaling, allocating more compute via longer token budgets to improve problem-solving. Identifying traces that reliably lead to correct answers is a key step toward improving the reliability and efficiency of these models. In this work, we propose *Latent-Space Metrics* that track the shifts in internal representations during the generation of intermediate reasoning tokens. We introduce a set of trajectory metrics that quantify both the magnitude of hidden-state changes and the geometry of their trajectories along the reasoning trace. We show that metrics tracking the model's internal states, rather than its output tokens, can serve as strong predictors of final answer accuracy. Our results demonstrate that they consistently distinguish correct from incorrect traces across models and reasoning domains. Moreover, we show that they enable more effective and efficient test-time scaling strategies, reducing token usage by up to 70% while preserving and even improving accuracy by 2.6% on average.

1 Introduction

Recent advances in large language models (LLMs) have shown that complex reasoning tasks can be solved more effectively by scaling computing during inference to generate longer and multiple chains-of-thought (*reasoning traces*) [7, 1, 14, 23]. However, not all traces are equally valuable: while some contain productive steps that lead to correct answers, others may deviate into unproductive paths such as overthinking, failing to reach a correct strategy, or reasoning inconsistently [e.g. 17, 2, 20]. Identifying which traces are likely to succeed is critical: it enables more reliable prediction of correct answers, improves efficiency by reducing wasted computation on unproductive paths, and can provide feedback signals that lead to better model training. By understanding which reasoning processes are effective, we can guide models to reinforce productive strategies and avoid erratic ones.

A growing body of work has attempted to characterize the quality of reasoning traces based on their surface form in natural language, a strategy that often requires costly annotation methods [12, 4]. In addition, natural language traces may not faithfully reflect the underlying strategies that models use [3, 19], and in some cases, models are trained to produce only latent embeddings rather than explicit text [8]. This suggests that language alone may be an unreliable proxy for evaluating the quality of the reasoning process.

An alternative perspective is to examine the model's latent space—i.e., the patterns of activity in its internal representations. Prior work has demonstrated that probing these representational spaces can reveal informative signals about safety [25], learning processes [10], and overall performance [21] of LLMs. On this basis, we hypothesize that the evolution of internal representations during the generation of intermediate reasoning tokens may also carry predictive signals about the accuracy of the final answer. Concretely, we ask:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

- Can latent-space metrics indicate the quality of a reasoning trace in terms of its likelihood of producing a correct final answer?
- Can these internal signals be harnessed to improve the efficiency and accuracy of reasoning models across domains?

To address these questions, we introduce a set of representational metrics that capture the temporal dynamics of hidden states within tokens in a reasoning trace. Specifically, we quantify how strongly the generation of a reasoning trace updates the latent space of a model, how much the representational trajectory shifts during reasoning, and how well these changes remain aligned. We apply these metrics to a range of models (DeepSeek-R1-Distill-Qwen-14B, Phi4 Reasoning Plus, and Qwen3-14B) across diverse reasoning domains, spanning science, math, and optimization problems. We show that they (1) reliably and significantly distinguish between traces leading to correct versus incorrect answers, and (2) can be leveraged at test time to achieve both higher efficiency and improved accuracy. Concretely, we introduce a test-time scaling strategy based on latent-space metrics that yields up to a 70% reduction in token usage compared to majority voting, along with 2.6% average improvement in accuracy.

2 Related Work

A growing body of work focuses on quantifying the quality of reasoning traces. Some approaches perform fine-grained analyses of the surface form of the reasoning trace, introducing metrics that capture their factual and logical correctness, as well as their linguistic and semantic coherence [22, 6]. These methods typically require annotating or systematically extracting information from traces, relying on either human annotators or auxiliary expert models. Other approaches evaluate trace quality by measuring its causal influence on the final answer [15]. Such methods, however, are computationally expensive, as they often involve generating multiple inference passes, perturbing or ablating parts of the reasoning trace, and re-evaluating model outputs to estimate causal effects. By contrast, latent-space metrics can be computed directly at inference time, without a teacher model or repeated runs, making them substantially more efficient.

More closely related to our work, authors in [10] analyze how internal representational trajectories in LLMs straighten during training, and speculate about the implications for model performance. We shift attention instead to temporal metrics tailored to reasoning traces, and empirically test their ability to predict answer quality. In a complementary direction, [21] examines representational curvature across layers (a spatial perspective) and demonstrate its predictive value for answer accuracy in chain-of-thought problems. Our work differs by adopting a temporal perspective (across tokens) and focusing specifically on reasoning models and sample aggregation strategies.

3 Methods

3.1 Preliminaries

Given a problem, reasoning models generate an output sequence consisting of a reasoning trace followed by a final answer. The reasoning trace is often delimited by special "think" tokens, such that:

$$q_1, \ldots, q_n \texttt{<think>} t_1, \ldots, t_r \texttt{</think>} a_1, \ldots, a_m,$$

where $q_1, \ldots, q_n$ are the input question tokens, $t_1, \ldots, t_r$ are the reasoning trace tokens, and $a_1, \ldots, a_m$ are the final answer tokens.

For each position $r \in \{1, \ldots, R\}$ within the reasoning trace, the model produces a hidden state of activations at each layer $l \in \{1, \ldots, L\}$, denoted by $h_l^{(r)} \in \mathbb{R}^d$. These hidden states form a two-dimensional array of d-sized representations, indexed by layer and token position, and encode the **latent space** of the model during the reasoning trace.

3.2 Latent-Space Metrics

To characterize how a model's latent space evolves over the course of intermediate reasoning, we introduce trajectory metrics that quantify the magnitude and geometry of changes in the hidden states along the trace.

First, to enhance the signal robustness and reduce the dimensionality, we partition the reasoning trace into non-overlapping **reasoning segments** (n tokens = 500, see Appendix A) and, for each layer, average the token representations within each segment. This procedure smooths local fluctuations in token-level dynamics while preserving the overall trajectory of the reasoning process.

For each layer l, let $h_l^{(n)}$ denote the averaged hidden state during reasoning segment n, with $n = 1, \ldots, N$. We define the **reasoning drift vector** as:

$$u_l = h_l^{(N)} - h_l^{(1)},$$

and measure the contribution of each reasoning segment by computing an **update vector** as:

$$v_l^{(n)} = h_l^{(n)} - h_l^{(n-1)}, \quad n = 2, \ldots, N$$

Each metric is computed per layer and then averaged across all layers to obtain a single scalar per reasoning trace.

Net Representational Change. A core question is whether intermediate reasoning substantially changes the model's internal representations. If reasoning has a real effect, the hidden states should be meaningfully updated after the trace. To assess this, we measure the overall magnitude of representational change the latent space undergoes from the first to the last reasoning segment. Concretely, we compute the norm of the reasoning drift vector, averaged across all layers and normalized by the number of segments:

$$\textsc{NetChange} = \frac{1}{L} \sum_{l \in L} \frac{\|u_l\|_2}{N}$$

Cumulative Representational Change. While NETCHANGE measures the overall representational change between the initial and final reasoning segments, it does not reflect how much fluctuation occurs along the trajectory itself, and cannot distinguish whether the trajectory was direct or highly varying. Excess cumulative changes in the latent space could indicate that the model is struggling to find the correct solution for the reasoning problem. To understand the total amount of movement the representations undergo during reasoning, we instead sum the magnitude of the segment-wise changes across the entire trace, and average across layers:

$$\textsc{CumulativeChange} = \frac{1}{L} \sum_{l \in L} \sum_{n=2}^{N} \|v_l^{(n)}\|_2$$

Aligned Representational Change. Beyond measuring how far representations move during reasoning, an important question is whether these movements advance toward the final state representation, rather than drifting sideways or reversing direction. We hypothesize that if reasoning is effective, the sequence of updates should align with the overall representational shift from the beginning to the end of the trace. To capture this property, we compute the cosine similarity between each update vector and the reasoning drift vector, averaging across segments and layers:

$$\textsc{AlignedChange} = \frac{1}{L} \sum_{l \in L} \frac{1}{N-1} \sum_{n=2}^{N} \frac{\langle v_l^{(n)}, u_l \rangle}{\|v_l^{(n)}\|_2 \, \|u_l\|_2}.$$

3.3 Models and Datasets

We evaluate three open-source reasoning models of 14B parameters: Deepseek-R1-Distill-Qwen-14B (R1-D) [7], Phi-4 Reasoning Plus Model (PHI4R+) [1], and Qwen3-14B (QWEN3) [23]. Our study probes these models across three distinct reasoning domains:

- *Scientific reasoning*, measured using the *GPQA Diamond* benchmark, which comprises 198 graduate-level multiple-choice questions in biology, chemistry, and physics [16].
- *Mathematical reasoning*, evaluated on *AIME 2025*, a 30-problem set from the American Invitational Mathematics Examination [11].
- *Algorithmic reasoning*, assessed with a stratified subsample ($n = 180$) of the *TSP* benchmark, consisting of path-optimization problems across varying levels of difficulty [5].

For our experiments, we sampled 5 answers to each of the problems in these datasets.

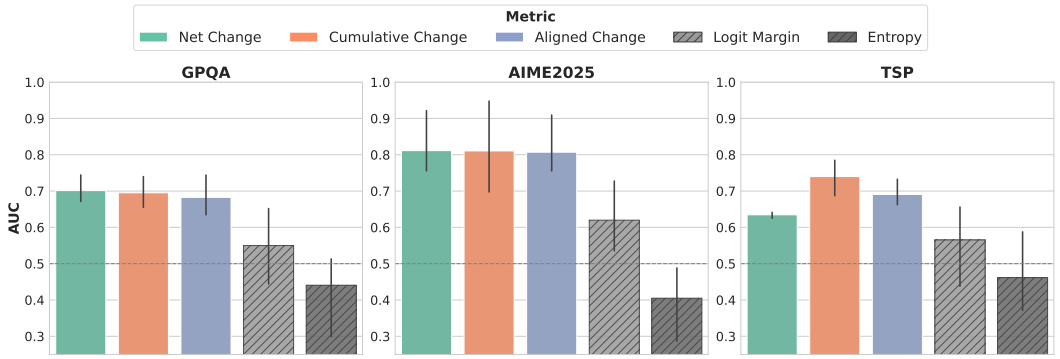

Figure 1: Area under the ROC curve (AUC) for distinguishing correct from incorrect predictions using latent-space and output-based metrics across datasets. Higher values indicate better discriminative power. Latent-space metrics consistently achieve significant and stronger discrimination than output-based metrics, highlighting their reliability for predicting correctness across datasets. For comparability, Cumulative Representational Change was sign-reversed so that larger values correspond to stronger discrimination. Error bars denote variability across models. The dashed line indicates chance-level performance.

4 Results

4.1 Latent-space metrics are predictive of solution accuracy

First, we study whether the defined latent space metrics are predictive of the accuracy of a model generated solution. For each latent-space metric, we quantified its ability to discriminate between correct and incorrect reasoning trajectories using the area under the receiver operating characteristic curve (AUC). We compute this metric across all 5 sampled solutions and problems in our datasets. To assess whether latent-space metrics provide stronger signals than output distribution–based measures commonly used as estimates of model confidence and uncertainty [24], we also compute the AUC of: (1) the *Logit Margin*, defined as the difference between the logits of the top-1 and top-2 predicted tokens when generating the final answer; and (2) the *Entropy* of the next-token distribution at the final answer prediction (see Appendix C for more details).

As Figure 1 shows, all latent-space metrics significantly distinguish between reasoning paths that lead to accurate versus inaccurate final answers (see also Appendix D). Across models and datasets, their AUCs remain consistently above chance, indicating reliable predictive power (Net Change mean AUC $= 0.71 \pm 0.09$; Cumulative Change $= 0.74 \pm 0.09$; Aligned Change $= 0.73 \pm 0.08$). In contrast, the output-distribution–based metrics are substantially weaker and less consistent across datasets and models, with performance often close to or below chance level (Logit Margin $= 0.59 \pm 0.10$; Entropy $= 0.44 \pm 0.10$).

We found that Cumulative Change was negatively correlated with accuracy (Pearson's correlation coefficient $r = -.19$), suggesting that traces wandering farther through representation space were less likely to yield correct answers. In contrast, both Net and Aligned Change correlated positively with accuracy (Net Change: $r = .15$; Aligned Change: $r = .32$), indicating that traces with higher overall representational effect, and whose intermediate updates advanced more directly toward the final state, were more likely to succeed. Figure 2 illustrates these distributions for one model and dataset; full results for all models and datasets are included in Appendix B.

4.2 Latent-Space metrics can improve efficiency and performance in reasoning problems

Building on our previous finding that latent-space metrics strongly predict solution accuracy, we now investigate whether these metrics can guide more effective and efficient test-time scaling strategies. We compared our approach against two sample aggregation baselines: (1) majority vote, the standard answer-selection method in the literature [1, 7] and (2) shortest-answer selection, which chooses

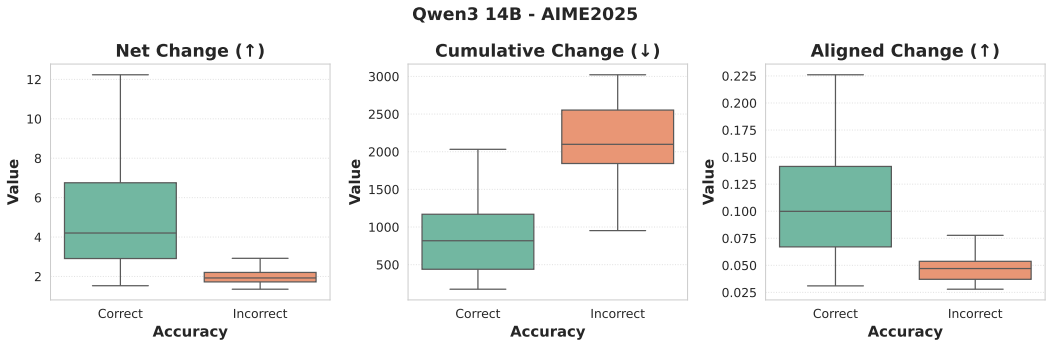

Figure 2: Latent-space metrics' distribution by accuracy for Qwen3-14B on the AIME 2025 dataset. Correct traces show larger Net/Aligned Change and smaller Cumulative Change than incorrect ones. This indicates that correct reasoning corresponds to larger, more directed representational shifts, while incorrect reasoning involves more wandering and less aligned trajectories. Equivalent plots for other models and datasets are provided in Appendix B.

out of five repeats the candidate with the fewest tokens, motivated by recent findings that shorter completions are strong signals of answer correctness [9, 18, 13].

For each datapoint, we sampled up to 5 candidate solutions from the model. If a solution's metric value exceeded the threshold, we immediately accepted the answer as the final prediction and stopped inference early. If none of the 5 sampled solutions for a datapoint crossed the threshold, we instead defaulted to majority voting across all candidates (see Figure 3). This design allows datapoints with strong internal signals to be resolved quickly with fewer samples, while datapoints with weaker signals fall back on the robustness of aggregation. We repeated this procedure independently for each latent-space metric under study.

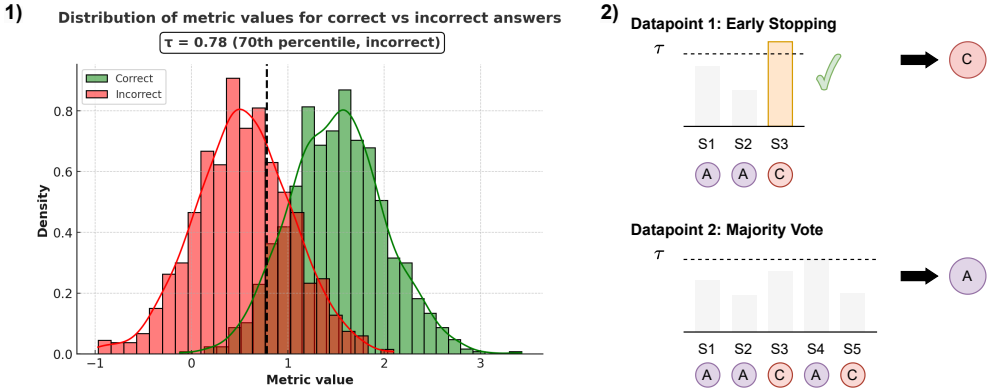

Figure 3: Illustration of our procedure for selecting reasoning traces. **(1)** Distribution of latent-space metric values for correct (green) and incorrect (red) answers. A threshold τ is chosen on a held-out calibration set as a quantile of the incorrect distribution (here, the 70th percentile). **(2)** At inference time, candidate solutions for a problem are evaluated sequentially. If a solution's metric value exceeds τ, it is immediately accepted as the final prediction (Datapoint 1). If no solution crosses the threshold, the final answer is chosen via majority voting across all candidates (Datapoint 2).

To determine thresholds, we used a cross-validation procedure (3 splits) with a calibration subset corresponding to 30% of the dataset for each split (see Appendix E). On the calibration subset, we generated candidate thresholds by taking quantiles of the metric values for incorrect solutions. In other words, each threshold corresponds to a cutoff such that a fixed proportion of incorrect

Table 1: Accuracy and efficiency using latent-space metrics. Baselines are Maj@5 (majority vote across 5 samples) and Shortest@5 (shortest of 5 samples [9, 18, 13]). Latent-Space (LS) strategies select the first answer candidate that reaches the threshold, otherwise default to majority vote. We report accuracy (%) and the average number of samples used per datapoint. Numbers in () show difference in accuracy and % reduction in token usage wrt majority vote baseline. Numbers in **bold** and **bold** indicate best and second-best results (highest accuracy/lowest token usage) in each group. ✓ indicates cases where the average number of samples required was reduced by at least half.

Model	Strategy	GPQA		AIME2025		TSP	
		Acc. (avg % / ΔAcc)	Samples (avg / ΔTok %)	Acc. (avg % / ΔAcc)	Samples (avg / ΔTok %)	Acc. (avg % / ΔAcc)	Samples (avg / ΔTok %)
R1-D	Maj@5	59.90	5.00	56.67	5.00	27.50	5.00
	Shortest@5	60.91 (+1.0)	5.00 (0)	50.00 (-6.7)	5.00 (0)	28.75 (+1.3)	5.00 (0)
	LS – Net	61.10 (+1.2)	**1.69** (+53.9) ✓	**61.90** (+5.2)	**1.22** (+68.7) ✓	28.60 (+1.1)	**1.43** (+70.6) ✓
	LS – Cumulative	**62.10** (+2.2)	1.88 (+48.1) ✓	58.70 (+2.0)	2.56 (+29.9)	**30.90** (+3.4)	**1.61** (+66.4) ✓
	LS – Aligned	61.10 (+1.2)	**1.58** (+57.0) ✓	**60.30** (+3.6)	**1.43** (+61.3) ✓	29.50 (+2.0)	2.08 (+57.2) ✓
	LS – Combined	**61.80** (+1.9)	1.89 (+47.3) ✓	**61.90** (+5.2)	2.06 (+43.9) ✓	**30.10** (+2.6)	**1.43** (+70.3) ✓
Phi4R+	Maj@5	**70.20**	5.00	80.00	5.00	41.25	5.00
	Shortest@5	69.19 (-1.0)	5.00 (0)	70.00 (-10.0)	5.00 (0)	38.75 (-2.5)	5.00 (0)
	LS – Net	68.80 (-1.4)	**2.97** (+20.2) ✓	79.40 (-0.6)	**2.19** (+41.1) ✓	42.30 (+1.1)	**1.59** (+67.2) ✓
	LS – Cumulative	**69.60** (-0.6)	**2.99** (+18.9) ✓	81.00 (+1.0)	**2.43** (+32.1) ✓	**44.40** (+3.1)	2.63 (+42.2) ✓
	LS – Aligned	**69.60** (-0.6)	3.40 (+14.5) ✓	**82.50** (+2.5)	2.51 (+30.8) ✓	**44.10** (+2.9)	**1.96** (+58.7) ✓
	LS – Combined	**69.60** (-0.6)	3.28 (+16.3) ✓	**82.60** (+2.6)	2.54 (+28.7) ✓	43.80 (+2.6)	2.30 (+50.5) ✓
Qwen3	Maj@5	63.96	5.00	70.00	5.00	36.25	5.00
	Shortest@5	**64.47** (+0.5)	5.00 (0)	80.00 (+10.0)	5.00 (0)	30.63 (-5.6)	5.00 (0)
	LS – Net	63.70 (-0.3)	**1.42** (+63.9) ✓	79.40 (+9.4)	1.60 (+57.3) ✓	35.40 (-0.9)	3.18 (+34.0)
	LS – Cumulative	63.30 (-0.7)	2.25 (+41.3) ✓	**84.10** (+14.1)	2.03 (+43.2) ✓	**36.30** (+0.1)	**1.64** (+65.5) ✓
	LS – Aligned	**64.20** (+0.2)	1.75 (+52.0) ✓	**80.90** (+10.9)	**1.59** (+58.2) ✓	**37.80** (+1.6)	**2.08** (+56.4) ✓
	LS – Combined	63.70 (-0.3)	**1.70** (+53.4) ✓	**80.90** (+10.9)	**1.49** (+60.3) ✓	36.00 (-0.3)	2.46 (+48.4) ✓

datapoints fall above or below it. For each candidate threshold, performance on the calibration subset was computed using the same inference rule as above (early accept on threshold crossing; otherwise majority vote). The threshold with the best calibration performance was then evaluated on the remaining 70% of the data. Metrics were reported in terms of both accuracy (the proportion of datapoints answered correctly) and efficiency (the percentage of reasoning tokens consumed relative to the full inference process, and the average number of samples needed). Results shown are averages over the cross-validation splits.

In addition to exploring each metric separately, we built a **combined latent-space score**. For each dataset, we quantified the predictive utility of each metric by calculating its absolute Pearson correlation with accuracy on 10% of the dataset (see details in Appendix F). Each metric was directionally aligned such that higher values consistently indicated better performance (i.e. cumulative representational change was sign-inverted). These correlations were normalized to form weights that sum to one, yielding an interpretable distribution of relative importance across metrics. The combined latent-space score for each sampled solution was then computed as a weighted sum of its metric values, where metrics more strongly aligned with accuracy on the calibration set contributed more heavily (see Appendix F for weights used).

As Table 1 shows, across datasets and models, latent-space metrics enable substantial compute savings while improving or preserving accuracy. Although the magnitude of improvements varies across models and datasets, compared to majority vote, latent-space strategies exhibit:

- *Sample savings*: the number of required samples is reduced on average by **58%** (range: 32–76%).
- *Token savings*: as a consequence of sample savings, token usage (and thereby inference cost) is reduced on average by **48%** (range: 14–70%).
- *Accuracy improvements*: accuracy increases on average by **2.64%** (range: -1.4–14.10%), in contrast to the shortest-length heuristic baseline, which reduces accuracy on average by 1.44%.

5 Conclusions

Our work introduced a set of metrics that track the temporal evolution of reasoning traces in latent space, and demonstrate that they carry strong predictive signals for final-answer correctness. Leveraging these signals during test-time scaling enables both accuracy gains and compute efficiency across models and reasoning domains.

We limited our analysis to widely used and standard reasoning model families. However, our approach is model-agnostic: it can be extended to training variants (e.g., SFT) and architectural variants that employ soft-token reasoning traces. Beyond inference-time use, these metrics provide actionable signals for fine-tuning and calibration, with the potential to steer models toward more reliable reasoning trajectories.

Limitations of our current study include the use of simple combined metric and thresholding strategies. Future work could automate their selection and fusion (e.g., learned ensembles or decision policies). In addition, we plan to explore a complementary set of metrics that not only characterize the reasoning evolution at the temporal level, but also capture how reasoning evolves across the model hierarchy and in semantic space.

References

[1] Marah Abdin et al. "Phi-4-reasoning technical report". In: *arXiv preprint arXiv:2504.21318* (2025).

[2] Xingyu Chen et al. "Do not think that much for 2+ 3=? on the overthinking of o1-like llms". In: *arXiv preprint arXiv:2412.21187* (2024).

[3] Yanda Chen et al. "Reasoning Models Don't Always Say What They Think". In: *arXiv preprint arXiv:2505.05410* (2025).

[4] Kanishk Gandhi et al. "Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars". In: *arXiv preprint arXiv:2503.01307* (2025).

[5] GeoMeterData. *NP-Hard Traveling Salesman Problem Instances*. `https://huggingface.co/datasets/GeoMeterData/nphard_tsp1`. Accessed: 2025-09-05. 2025.

[6] Olga Golovneva et al. "Roscoe: A suite of metrics for scoring step-by-step reasoning". In: *arXiv preprint arXiv:2212.07919* (2022).

[7] Daya Guo et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning". In: *arXiv preprint arXiv:2501.12948* (2025).

[8] Shibo Hao et al. "Training large language models to reason in a continuous latent space". In: *arXiv preprint arXiv:2412.06769* (2024).

[9] Michael Hassid et al. "Don't Overthink it. Preferring Shorter Thinking Chains for Improved LLM Reasoning". In: *arXiv preprint arXiv:2505.17813* (2025).

[10] Eghbal Hosseini and Evelina Fedorenko. "Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language." In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 43918–43930.

[11] Hugging Face Datasets. *AIME 2025 Dataset (American Invitational Mathematics Examination 2025)*. `https://huggingface.co/datasets/opencompass/AIME2025`. Accessed: 2025-09-05. 2025.

[12] Seongyun Lee et al. "The CoT Encyclopedia: Analyzing, Predicting, and Controlling how a Reasoning Model will Think". In: *arXiv preprint arXiv:2505.10185* (2025).

[13] Sara Vera Marjanović et al. "DeepSeek-R1 Thoughtology: Let's think about LLM Reasoning". In: *arXiv preprint arXiv:2504.07128* (2025).

[14] OpenAI. *Learning to Reason with LLMs*. `https://openai.com/index/learning-to-reason-with-llms`. Accessed: 2025-09-09. 2024.

[15] Archiki Prasad et al. "Receval: Evaluating reasoning chains via correctness and informativeness". In: *arXiv preprint arXiv:2304.10703* (2023).

[16] David Rein et al. "Gpqa: A graduate-level google-proof q&a benchmark". In: *First Conference on Language Modeling*. 2024.

[17] Parshin Shojaee et al. "The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity". In: *arXiv preprint arXiv:2506.06941* (2025).

[18] Vaishnavi Shrivastava et al. “Sample More to Think Less: Group Filtered Policy Optimization for Concise Reasoning”. In: *arXiv preprint arXiv:2508.09726* (2025).

[19] Kaya Stechly et al. “Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens”. In: *arXiv preprint arXiv:2505.13775* (2025).

[20] Yiyou Sun et al. “OMEGA: Can LLMs Reason Outside the Box in Math? Evaluating Exploratory, Compositional, and Transformative Generalization”. In: *arXiv preprint arXiv:2506.18880* (2025).

[21] Yiming Wang et al. “Latent space chain-of-embedding enables output-free llm self-evaluation”. In: *arXiv preprint arXiv:2410.13640* (2024).

[22] Juncheng Wu et al. “Knowledge or Reasoning? A Close Look at How LLMs Think Across Domains”. In: *arXiv preprint arXiv:2506.02126* (2025).

[23] An Yang et al. “Qwen3 technical report”. In: *arXiv preprint arXiv:2505.09388* (2025).

[24] Gal Yona, Amir Feder, and Itay Laish. “Useful confidence measures: Beyond the max score”. In: *arXiv preprint arXiv:2210.14070* (2022).

[25] Andy Zou et al. “Representation engineering: A top-down approach to ai transparency”. In: *arXiv preprint arXiv:2310.01405* (2023).

A Representational Averaging

To enhance the signal robustness and reduce the dimensionality of the reasoning trace, we partition it into non-overlapping **reasoning segments** of size 500. For each layer, we then average the token representations within each segment. We found that this procedure preserves the overall trajectory of the reasoning process.

The choice of 500 tokens was guided by the average answer lengths across datasets. The dataset with the shortest responses still had an average of 5,000 tokens per answer. Setting the window to 500 tokens therefore ensures that, on average, we obtain at least 10 measurement points per answer in this dataset, and proportionally more in the others.

We also compared this strategy to defining segments by newline tokens. However, segment sizes varied substantially across models under with this approach, making it less consistent than fixed-length segmentation.

B Latent-Space Metrics

To investigate how latent-space dynamics relate to model performance, we compare distributions of three change-based metrics—net change, cumulative change, and aligned change—conditioned on whether a model's prediction was correct or incorrect. Figures 4–6 present box plots of these metrics across datasets and model families. These visualizations allow us to assess whether systematic differences in latent-space evolution are associated with answer correctness, and whether such effects are consistent across evaluation settings.

Across all three metrics, consistent patterns emerge. Net change values (Figure 4) are generally higher for correct responses than for incorrect responses, suggesting that successful reasoning is associated with overall larger representational drifts. In contrast, cumulative change values (Figure 5) are often larger for incorrect responses, indicating that when models answer incorrectly, their latent trajectories tend to involve more movement through representational space, potentially reflecting less stable reasoning. Finally, aligned change values (Figure 6) are again higher for correct responses, implying that effective reasoning requires updates that advance more directly towards the final state.

Taken together, these results suggest that correct predictions are characterized by a larger overall representational shift, accompanied by trajectories that are more directionally consistent, whereas incorrect predictions tend to involve longer, less aligned paths through latent space, reflecting noisier and less stable reasoning trajectories. This pattern holds across models and datasets, indicating that latent-space metrics provide complementary and reliable signals of reasoning quality.

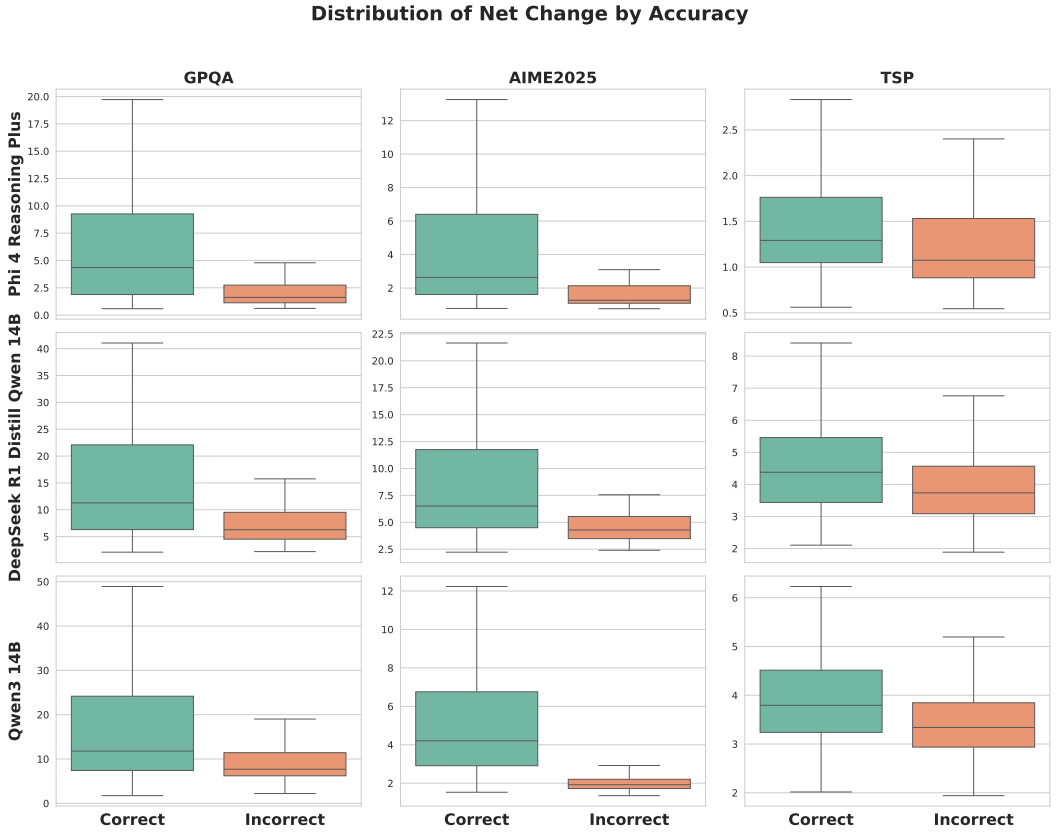

Figure 4: Distribution of net change metric values by accuracy. Net Change values are generally higher for correct responses than for incorrect responses, suggesting that successful reasoning is associated with overall larger representational drifts that may be a sign of deeper reasoning.

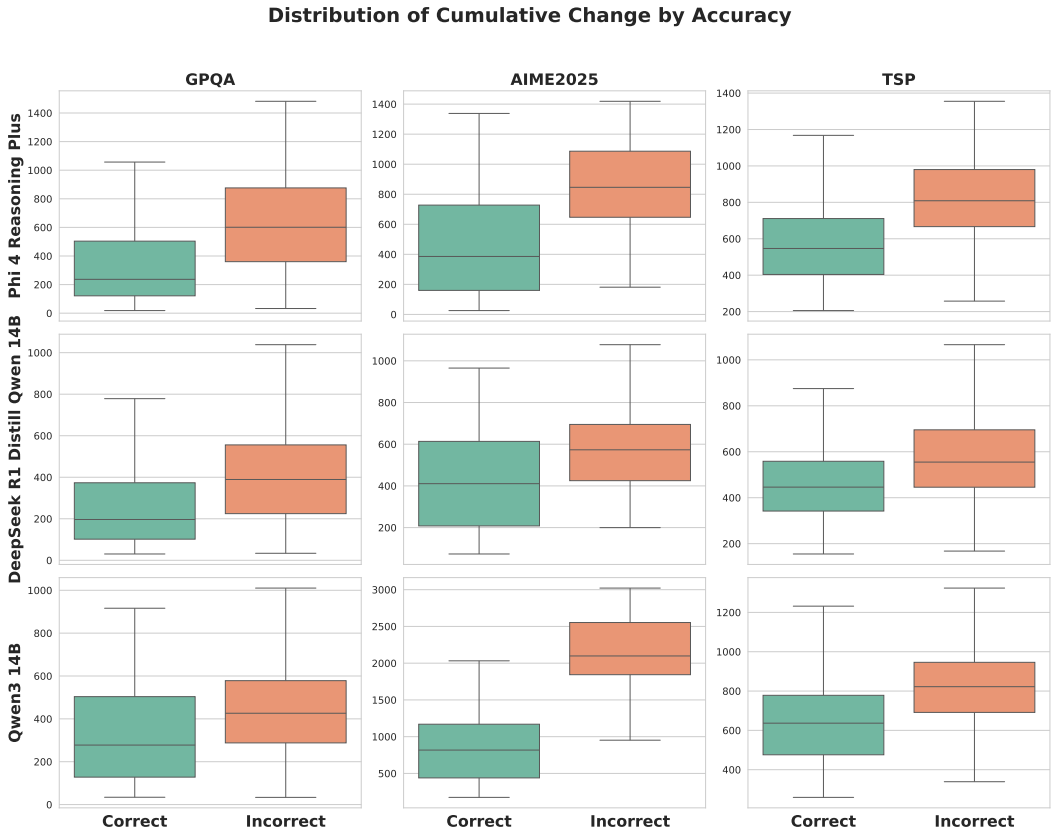

Figure 5: Distribution of cumulative change metric values by accuracy. Cumulative Change values are often larger for incorrect responses, indicating that when models answer incorrectly, their latent trajectories tend to involve more movement through representational space, potentially reflecting less stable reasoning.

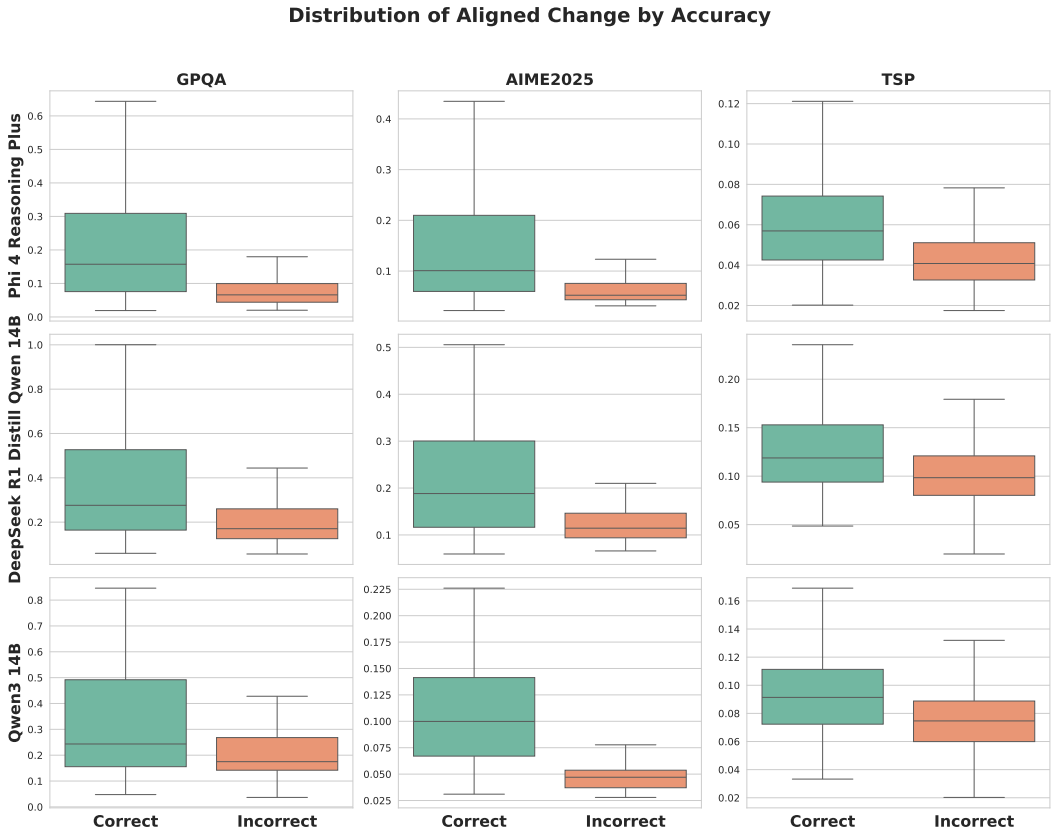

Figure 6: Distribution of aligned change metric values by accuracy. Averaged values are higher for correct responses, implying that effective reasoning involves intermediate representational updates that advance more directly towards the final reasoning state.

C Comparison with Output Distribution Metrics

We compared the predictive power of latent-space signals with that of the model's output distribution. Output distribution metrics directly encode the model's uncertainty and confidence over possible continuations, and can be used as a proxy for final answer reliability.

To access the output distribution, we prompted the model to produce a final answer following its reasoning trace. Specifically, we used prompts of the form:

$$q_1, \ldots, q_n \texttt{ <think> } t_1, \ldots, t_r \texttt{ </think> Final answer:}$$

We then examined the distribution over the next token. To quantify its informativeness, we computed two measures:

- **Logit Margin**: the difference between the logits of the two most probable tokens.
- **Entropy**: the entropy of the next-token distribution.

D Threshold Performance

We computed the accuracy of the samples whose metric values exceeded thresholds corresponding to quantiles from the 10th to the 90th percentile, in increments of 10. As shown in Figure 7, accuracy increases consistently with higher quantiles across all models and datasets. Datapoints that lie above higher latent-space quantiles are increasingly likely to be correct. This indicates that latent-space metrics are predictive of reasoning quality in a general way, not tied to any single model architecture or benchmark. In several cases, near-ceiling accuracy is achieved at the highest quantiles, demonstrating that filtering traces by high latent-space thresholds isolates a subset of datapoints that are almost always correct.

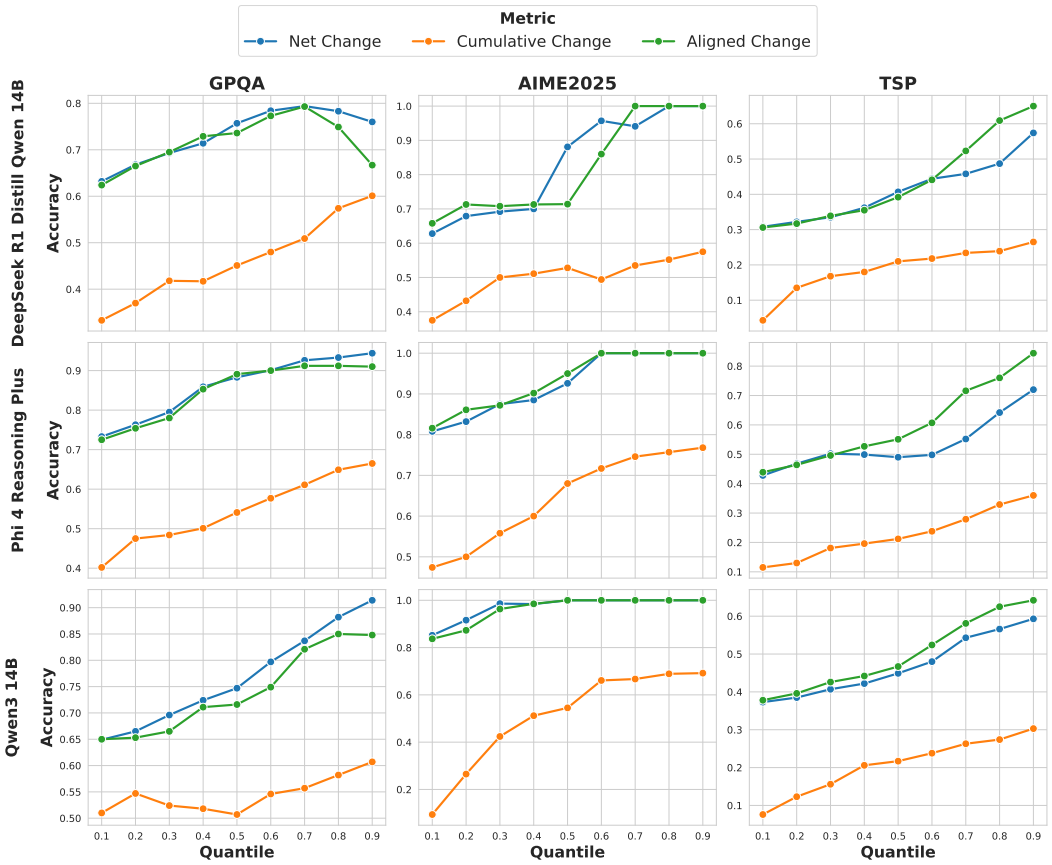

Figure 7: Accuracy of datapoints above thresholds defined over a range of quantiles. Given that Cumulative Change is negatively correlated with accuracy, we inverted the quantile selection (i.e., replacing q with $1 - q$) so that higher quantiles consistently correspond to higher expected accuracy. For all metrics, accuracy increases consistently with higher quantiles across datasets and models. This evidences that these metrics are predictive of answer quality.

E Cross-validated thresholding

Data splits and preprocessing. We use `ShuffleSplit` cross-validation with $n\_\text{splits} = 3$, a calibration–test partition of 30/70 per split, and a fixed random seed.

Candidate thresholds. Within each calibration fold and for each metric, we restrict to the calibration set's *incorrect* solutions:

$$\texttt{false} = \{\text{rows in train with } \texttt{accuracy} = \text{False}\}.$$

If $|\texttt{false}| < \texttt{min\_incorrect}$ $(= 15)$ or the metric has no valid values in `false`, we fall back to the calibration median:

$$\texttt{best\_thr} \leftarrow \text{median}(\texttt{calibration}[\text{metric}]).$$

Otherwise, we build a quantile grid:

$$Q = \{0.20, 0.21, \ldots, 0.99\},$$

and form the candidate threshold set

$$\mathcal{T} = \{\text{quantile}(\texttt{false}[\text{metric}], q) \ : \ q \in Q\}.$$

Scoring a candidate threshold. For each $t \in \mathcal{T}$, we partition calibration datapoints by applying the threshold to *sequential* candidates:

- **Threshold group** $\mathcal{D}_{\text{thr}}$: datapoints for which at least one candidate crosses the threshold. For these, we accept the *first* crossing candidate and compute its accuracy a_{thr} as the mean of `accuracy` over the accepted rows.
- **Fallback group** $\mathcal{D}_{\text{mv}}$: all remaining datapoints. We resolve them with majority vote (MV) over their candidates on the training set and obtain a_{mv}.

Let $n_{\text{thr}} = |\mathcal{D}_{\text{thr}}|$ and $n_{\text{mv}} = |\mathcal{D}_{\text{mv}}|$. The combined calibration accuracy for threshold t is the datapoint-weighted average:

$$A_{\text{calibration}}(t) = \frac{a_{\text{thr}} \cdot n_{\text{thr}} + a_{\text{mv}} \cdot n_{\text{mv}}}{\max(1,\ n_{\text{thr}} + n_{\text{mv}})}.$$

Choosing the threshold. We rank all $t \in \mathcal{T}$ by $A_{\text{calibration}}(t)$, take the top-2 thresholds, and set

$$\texttt{best\_thr} \ = \ \text{median}\Big(\arg\text{top}\_t \in \mathcal{T}^{\,2}\, A_{\text{calibration}}(t)\Big).$$

If $\mathcal{T}$ is empty, we again fall back to the calibration median of the metric.

Direction of comparison. For the *Cumulative Change* metric, lower values indicate stronger signals; accordingly, we apply the threshold as $\text{metric} \le t$. For all other metrics, higher is better and the threshold is applied as $\text{metric} \ge t$.

F Combined Latent-Space Score

Table 2 reports the weights for each latent-space metric, estimated from a 10% subset of the dataset.

Table 2: Metric weights for Combined Latent Space score.

Model	**Dataset**	Net Change	Cumulative Change	Aligned Change
DeepSeek R1 Distill Qwen 14B	GPQA	0.35	0.40	0.25
Phi 4 Reasoning Plus	GPQA	0.30	0.38	0.31
Qwen3 14B	GPQA	0.43	0.25	0.32
DeepSeek R1 Distill Qwen 14B	AIME2025	0.31	0.35	0.34
Phi 4 Reasoning Plus	AIME2025	0.26	0.45	0.29
Qwen3 14B	AIME2025	0.30	0.43	0.28
DeepSeek R1 Distill Qwen 14B	TSP	0.24	0.39	0.37
Phi 4 Reasoning Plus	TSP	0.20	0.38	0.42
Qwen3 14B	TSP	0.19	0.43	0.37