Intention Collapse: Intention-Level Metrics for Reasoning in Language Models

Anonymous ACL submission

Abstract

Large language models are built from relatively simple computational machinery, yet their performance across tasks and prompting strategies remains difficult to predict in advance. In particular, reasoning interventions such as chain-of-thought (CoT) can substantially help in some settings while degrading performance in others. We study this gap through the lens of *Intention Collapse*, the transition from a prompt-conditioned pre-emission state to a discrete output sequence. We introduce three lightweight, model-agnostic diagnostics of this pre-emission state: intention entropy, effective dimensionality, and recoverability (probe AUROC for eventual correctness). Across a 3×3 study spanning three instruction-tuned model families (Mistral-7B, LLaMA-3.1-8B, Qwen-2.5-7B) and three reasoning benchmarks (GSM8K, ARC-Challenge, AQUA-RAT), we compare direct answering, CoT, and a verbosity control. We find that CoT is not universally beneficial: it yields large gains on GSM8K but degradations on ARC-Challenge in two of three model families, with a near-null effect on Mistral-7B, with mixed effects on AQUA-RAT. We further find that CoT induces distinct pre-emission entropy regimes across model families, and that recoverable internal signal can remain above chance even when final answer accuracy declines, especially in multiple-choice settings. These results suggest that part of the challenge in understanding LLM reasoning is not only how models compute, but how reliably informative pre-emission states are converted into valid final decisions. Intention Collapse provides a simple empirical framework for studying that boundary.

1 Introduction

Large language models (LLMs) are built from relatively simple computational primitives, yet their behavior across tasks and prompting strategies remains difficult to forecast in advance. This challenge is especially visible in reasoning interventions such as chain-of-thought (CoT), which can substantially improve performance in some settings while degrading it in others (Wei et al., 2022). We still lack lightweight empirical tools for characterizing when such interventions help, when they hurt, and what changes internally before a model commits to a final output.

We study this gap through the lens of *Intention Collapse*: the transition from a prompt-conditioned pre-emission state to a discrete output sequence. We formalize an intention state I as the latent representation immediately before the first token is emitted, and a collapse operator κ_θ that maps I into language. Because this mapping is many-to-one and lossy, information present in I may fail to survive into a valid final answer.

This perspective is motivated by prior work on reasoning interventions (Wei et al., 2022; Zelikman et al., 2022; Lightman et al., 2023; Sun et al., 2019; Madaan et al., 2023) and by related studies on latent knowledge elicitation, pre-generation adequacy, and uncertainty estimation (Burns et al., 2022; Yoon et al., 2025; Kuhn et al., 2023). To operationalize the framework, we introduce three lightweight, model-agnostic diagnostics of the pre-emission state: **intention entropy** $H_{\text{int}}(I)$, **effective dimensionality** $d_{\text{eff}}(I)$, and **recoverability** $\text{Recov}(I; Z)$, defined here as probe AUROC for eventual correctness.

We evaluate these diagnostics in a controlled 3×3 study spanning three instruction-tuned model families (Mistral-7B-Instruct, Llama-3.1-8B-Instruct, and Qwen-2.5-7B-Instruct), three reasoning benchmarks (GSM8K, ARC-Challenge, and AQUA-RAT), and three inference regimes (direct answering, CoT, and a Babble control). CoT is not a universal improvement: it yields large gains on GSM8K, degradations on ARC-Challenge in two of three model families, with a near-null effect on Mistral-7B, and mixed effects on AQUA-RAT.

We further find that CoT induces family-specific entropy shifts, that response format is a major axis of variation, and that recoverable internal signal can remain above chance even when final accuracy declines, especially in multiple-choice settings.

These findings support a narrower but useful claim: understanding LLM reasoning requires analyzing not only how models compute, but how reliably informative pre-emission states are converted into valid task-specific decisions. In this sense, Intention Collapse is a lightweight empirical framework for studying the boundary between internal representation and final output, rather than a complete theory of reasoning.

2 Operationalizing Intention Collapse

To transform *Intention Collapse* into a measurement, we require (i) a precise definition of the pre-emission state I, (ii) a reproducible extraction point, and (iii) lightweight diagnostics that quantify uncertainty, geometry, and recoverable task-relevant signal in I. Throughout, we treat I as a strictly pre-generation object: it is extracted after prompt conditioning and before the first output token is selected.

2.1 Intention State

For a transformer-based language model with L layers processing a prompt $x_{1:S}$, we define the *intention state* I as the collection of hidden representations at the first decoding step, immediately before the first output token is emitted:

$$I \triangleq \{h_{\text{pre}}^{(\ell)}\}_{\ell \in \mathcal{L}} \in \mathbb{R}^{|\mathcal{L}| \times d}, \tag{1}$$

where $h_{\text{pre}}^{(\ell)} \in \mathbb{R}^d$ denotes the hidden state at layer ℓ at the final prompt position, and $\mathcal{L} \subseteq \{1, \ldots, L\}$ is a fixed pre-specified subset of layers drawn from the upper third of each network (layers 27–31 for Mistral-7B, layers 28–32 for LLaMA-3.1-8B, and layers 24–28 for Qwen-2.5-7B, layer indices are 1-based throughout), used consistently across all conditions within each model. We focus on the upper third because prior work on transformer interpretability has repeatedly shown that late layers carry task-relevant and decision-relevant information closest to the readout boundary, while early layers are dominated by token-level and syntactic features that are less informative about eventual answer correctness (Elhage et al., 2021; Burns et al., 2022). The specific five-layer band within each model's upper third was selected once before any analysis, based on the late-layer convention in latent-knowledge probing, and was held constant across all regimes, benchmarks, and metrics to prevent layer-selection from acting as a hidden degree of freedom. A full per-layer sensitivity sweep across all 27 cells is beyond the scope of this paper and is left to future work; the same fixed layer subset is used throughout the appendix robustness analyses.

Operationally, the model first processes the full prompt, including any regime-specific instruction (e.g., direct answering, CoT, or Babble). We then record hidden states at the last prompt position across selected layers, before any discrete token commitment occurs. This definition makes I strictly pre-collapse: it captures the prompt-conditioned internal state immediately before generation begins. As in related work on latent knowledge elicitation (Burns et al., 2022), pre-generation adequacy (Yoon et al., 2025), and uncertainty estimation (Kuhn et al., 2023), the goal is to study what is already present in the representation prior to output.

2.2 Intention Metrics

We introduce three complementary, model-agnostic diagnostics of the pre-emission state.

Intention entropy. We define intention entropy as the Shannon entropy of the next-token distribution at the commitment boundary:

$$H_{\text{int}}(I) \triangleq -\sum_{v \in \mathcal{V}} p_\theta(y_1 = v \mid x_{1:S}) \log_2 p_\theta(y_1 = v \mid x_{1:S}), \tag{2}$$

where $\mathcal{V}$ is the model vocabulary. Lower entropy indicates a more concentrated continuation distribution. In this paper, $H_{\text{int}}(I)$ is computed over the full vocabulary and is interpreted as a measure of local generative dispersion at the pre-emission boundary, rather than as a calibrated uncertainty measure over task-specific answer choices. This distinction is especially important in multiple-choice settings, where full-vocabulary entropy is only loosely related to uncertainty over the valid answer set.

Effective dimensionality. To characterize the geometry of I, we compute per-layer effective dimensionality using the participation ratio of the PCA eigenvalue spectrum:

$$d_{\text{eff}}^{(\ell)} \triangleq \frac{\left(\sum_{j=1}^{d} \lambda_j^{(\ell)}\right)^2}{\sum_{j=1}^{d} \left(\lambda_j^{(\ell)}\right)^2}, \tag{3}$$

where $\{\lambda_j^{(\ell)}\}_{j=1}^{d}$ are the eigenvalues of the sample covariance matrix of hidden states at layer ℓ. We then aggregate across layers using a variance-weighted average:

$$d_{\text{eff}}^{\text{agg}} \triangleq \sum_{\ell \in \mathcal{L}} w_\ell \, d_{\text{eff}}^{(\ell)}, \qquad w_\ell = \frac{\operatorname{tr}(C^{(\ell)})}{\sum_{\ell' \in \mathcal{L}} \operatorname{tr}(C^{(\ell')})}. \tag{4}$$

Higher d_{eff} indicates that the pre-emission representation spans a broader effective subspace under this participation-ratio summary. We aggregate across layers with variance-based weights in order to summarize geometry without concatenating heterogeneous layer spaces into a single global PCA. We do not interpret d_{eff} as a direct measure of reasoning quality; rather, it serves as a complementary geometric diagnostic whose qualitative shifts are cross-checked in the appendix with alternative intrinsic-dimension estimators.

Recoverability. To estimate how much task-relevant signal remains linearly accessible before emission, we define recoverability as the performance of a linear probe trained on I to predict a target variable Z, here eventual answer correctness:

$$\text{Recov}(I; Z) \triangleq \text{AUROC}\big(f_\phi(I), Z\big), \tag{5}$$

where f_ϕ is an ℓ_2-regularized logistic regression classifier over concatenated hidden states. We report AUROC because it is robust to class imbalance. Probes are trained with item-level train/validation/test splits, and full details on regularization, sanity checks, and cross-regime transfer are provided in the appendix.

Together, these metrics quantify three complementary aspects of the pre-emission state: local uncertainty, representational geometry, and recoverable task-relevant information. They require no architectural modification and can be applied to standard transformer-based models.

2.3 Experimental Conditions

We evaluate intention metrics under three inference regimes while holding decoding fixed (greedy decoding, temperature $T = 0$) in order to isolate prompt effects: **Baseline** (direct answer instruction), **CoT** ("think step by step"), and **Babble** (length-oriented stream-of-consciousness control). CoT follows prior work on explicit intermediate reasoning (Wei et al., 2022), while Babble serves as a verbosity control that does not aim to solve the task.

In all three regimes, I is extracted at the same computational moment: after prompt processing and before the first generated token. The regimes differ only in the instructions embedded in the prompt and therefore in how prompt conditioning shapes the pre-emission state. This design cleanly separates changes attributable to pre-collapse conditioning from post-collapse generation dynamics.

We instantiate this framework in a controlled 3×3 design spanning three model families (Mistral-7B-Instruct, Llama-3.1-8B-Instruct, and Qwen-2.5-7B-Instruct), three reasoning bench- marks (GSM8K, ARC-Challenge, and AQUA-RAT), and the three inference regimes above, with $n = 200$ items per model–benchmark cell. This yields a balanced dense matrix for comparing how prompt strategy reshapes the state immediately before generation and whether those shifts align with downstream performance. While this sample size does not support strong claims about small effects or higher-order interactions, it is sufficient to expose several stable qualitative signatures across models, tasks, and regimes; we therefore emphasize repeated cross-condition patterns rather than isolated single-cell effects.

3 Experimental Study

The usefulness of *Intention Collapse* depends on whether pre-emission measurements reveal stable and interpretable structure across models, tasks, and prompting regimes. We therefore evaluate the framework in a controlled 3×3 study designed to test three questions: (i) how prompting strategies reshape the pre-emission intention state I, (ii) when such reshaping improves or degrades downstream accuracy, and (iii) whether I contains recoverable task-relevant signal that is not reliably expressed in the final output. Our focus is not on maximizing benchmark performance, but on characterizing the boundary between internal representation and final task-specific commitment.

3.1 Experimental Setup

We evaluate three instruction-tuned open-weight models—Mistral-7B-Instruct, Llama-3.1-8B-Instruct, and Qwen-2.5-7B-Instruct (Jiang et al., 2023; Grattafiori et al., 2024; Yang et al., 2024)—on three reasoning benchmarks spanning free-response and multiple-choice formats: GSM8K (Cobbe et al., 2021), ARC-Challenge, and AQUA-RAT. GSM8K provides a free-response

math setting; ARC-Challenge tests multiple-choice abstract reasoning; and AQUA-RAT provides a multiple-choice math setting that allows within-domain comparison of response format.

For each model–benchmark cell, we sample $n = 200$ items and evaluate three inference regimes: Baseline (direct answer), CoT ("think step by step"), and Babble (a length-oriented control unrelated to solving). Decoding is held fixed (greedy, $T = 0$) to isolate prompt-conditioning effects under a controlled commitment policy. Sampling-based decoding would introduce an additional source of stochastic variation at the commitment boundary, making it harder to attribute observed effects to changes in the pre-emission state itself. Accordingly, the results should be read as characterizing the pre-emission boundary under controlled deterministic decoding, rather than as a full evaluation of standard CoT practice under sampling-based strategies such as self-consistency. For multiple-choice tasks, we extract a single final option letter with a conservative parser; for GSM8K, we extract the final numeric answer using standard normalization. Extraction failures are counted as incorrect.

For each run, we extract the intention state I at the last prompt position immediately before the first output token is selected, and compute the three diagnostics introduced in Section 2: intention entropy $H_{\text{int}}(I)$, effective dimensionality $d_{\text{eff}}(I)$, and recoverability $\text{Recov}(I; Z)$. Recoverability is measured with ℓ_2-regularized logistic regression probes trained to predict eventual correctness, using item-level train/validation/test splits (60/20/20) and AUROC with bootstrap confidence intervals. Full details on parsing, compliance diagnostics, robustness checks, and probe controls are reported in the appendix.

3.2 Main Results

CoT is not universally beneficial. Across the 3×3 matrix, CoT increases mean accuracy from 34.2% to 47.3%, driven primarily by large gains on GSM8K across all three model families. However, CoT consistently degrades performance on ARC-Challenge and produces mixed effects on AQUA-RAT. This result is central to our framing: prompting strategies that appear to elicit more reasoning do not uniformly improve final task performance, especially across different response formats. To make the reporting more explicit, Appendix B.4 (Table 3) includes Wilson 95% confidence intervals for all per-cell accuracy estimates; these intervals

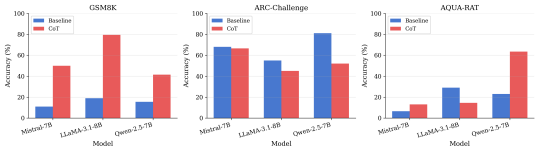

Figure 1: Accuracy across Baseline, CoT, and Babble for the 3×3 model–benchmark matrix ($n = 200$ per cell). CoT yields large gains on GSM8K across models, but degrades ARC-Challenge across all three model families, illustrating strong response-format dependence.

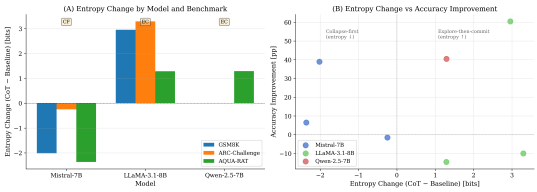

Figure 2: Entropy regime shifts under CoT, $\Delta H = H_{\text{int}}(\text{CoT}) - H_{\text{int}}(\text{Base})$, highlighting model-family heterogeneity. CoT does not induce a uniform pre-emission generative dispersion signature across models.

show that the major gains and degradations emphasized in the matrix are not small-sample artifacts, while Mistral/ARC-Challenge remains consistent with a near-null CoT effect.

CoT induces distinct entropy regimes across model families. To quantify how prompting reshapes the pre-emission state, we compare entropy under CoT and Baseline using $\Delta H = H_{\text{int}}(\text{CoT}) - H_{\text{int}}(\text{Base})$. We observe qualitatively different regimes across model families: Mistral tends to operate under lower pre-emission entropy in CoT ($\Delta H < 0$), while LLaMA tends to operate under higher pre-emission entropy ($\Delta H > 0$). These results indicate that CoT does not impose a universal internal signature, but instead interacts with family-specific representational dynamics. Qwen follows the same explore-then-commit pattern as LLaMA across all three benchmarks

Response format is a major axis of variation. A contrast appears within math-like tasks: GSM8K (free-response) benefits strongly from CoT, while AQUA-RAT (multiple-choice) shows mixed and model-dependent effects. Together with the ARC-Challenge degradations, this suggests that part of the difficulty is not only whether a model contains useful internal signal, but whether that signal can be converted into a valid discrete decision under the required format. For multiple-choice tasks, this includes a compliance axis: the model must not only reason correctly but also produce exactly one

Table 1: Compliance-conditioned accuracy for multiple-choice tasks. We report compliance rate with Wilson 95% confidence intervals; Acc | Comp. denotes accuracy among compliant generations only.

Model	Bench.	Regime	Compl.	[95% CI]	Acc. Comp.
Mistral	AQUA	Base	0.275	[0.218, 0.341]	0.236
Mistral	AQUA	CoT	0.525	[0.456, 0.593]	0.248
LLaMA	AQUA	Base	0.930	[0.886, 0.958]	0.312
LLaMA	AQUA	CoT	0.240	[0.186, 0.304]	0.604
LLaMA	ARC	Base	0.905	[0.856, 0.938]	0.608
LLaMA	ARC	CoT	0.610	[0.541, 0.675]	0.738
Qwen	AQUA	Base	0.780	[0.718, 0.832]	0.295
Qwen	AQUA	CoT	0.775	[0.712, 0.827]	0.819
Qwen	ARC	Base	1.000	[0.981, 1.000]	0.810
Qwen	ARC	CoT	1.000	[0.981, 1.000]	0.520

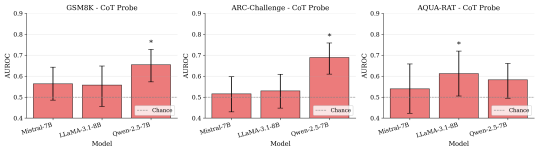

Figure 3: Probe AUROC predicting eventual correctness from the pre-emission state I under the CoT/enhanced condition. In a limited subset of settings, recoverability remains reliably above chance even when final answer accuracy is weak, indicating a dissociation between internal signal and successful final commitment.

valid option under the required interface. To quantify this effect directly, Table 1 decomposes performance into compliance rate and accuracy among compliant generations only. The resulting picture is mixed but informative. For LLaMA, CoT lowers parsed accuracy on both AQUA-RAT and ARC-Challenge while increasing accuracy conditional on compliance, indicating that a substantial portion of the degradation arises from format-commitment failures rather than loss of task-relevant internal signal. In contrast, Qwen/ARC-Challenge maintains perfect compliance in both regimes, so its CoT degradation cannot be attributed to formatting alone; and for Mistral/AQUA-RAT, CoT improves mainly through higher compliance while accuracy conditional on compliance changes only slightly. We therefore interpret compliance as an important but non-exhaustive contributor to multiple-choice performance differences under CoT: Table 1 shows that some degradations are strongly format-driven, while others persist even under perfect compliance and are therefore not reducible to formatting alone. This is consistent with a broader commitment-failure view of MCQA, in which performance depends not only on internal winner selection but also on the reliability with which that selection is bound to a valid output symbol at emission time. Appendix A.4 reports the parsing and compliance protocol.

Recoverable internal signal can persist even when final accuracy drops. Focusing on the CoT/enhanced condition used in the main comparison, linear probes trained on I achieve statistically reliable AUROC above chance in three cells where bootstrap CIs exclude 0.5: Qwen/ARC-Challenge (0.69, CI $[0.61, 0.76]$), Qwen/GSM8K (0.65, CI $[0.57, 0.73]$), and LLaMA/AQUA-RAT (0.61, CI $[0.51, 0.72]$). In the remaining CoT/enhanced cells, point estimates exceed 0.5 but CIs include chance; these effects should be treated as suggestive rather than conclusive at this sample size. Baseline probe AUROCs are reported separately in Appendix B.2 and should not be read as evidence for the same CoT-specific dissociation claim. Importantly, statistically reliable representational separability can co-occur with degraded behavioral performance: Qwen/ARC-Challenge achieves probe AUROC 0.69 despite a 29 pp accuracy drop under CoT. This dissociation supports a core claim of the intention-collapse perspective: task-relevant signal may be present in the pre-emission state without being reliably expressed in the final output.

Babble isolates verbosity from reasoning. Babble produces long outputs but does not reproduce the metric–performance patterns observed under CoT. This suggests that CoT effects are not explained by output length alone, but by changes in the pre-emission state I and in the reliability of final commitment. Taken together, these results show that CoT is not always beneficial: it helps on some tasks, hurts on others, and its effects vary across model families. In some cases, useful pre-emission signal remains present even when the final answer gets worse. This supports analyzing not only final outputs, but also how internal signal is converted into a final decision.

4 Related Work

Our framework connects three lines of work. First, studies on **latent knowledge and interpretability** show that language models often encode more in their activations than they express in outputs (Elhage et al., 2021; Conmy et al., 2023; Christiano and Xu, 2021; Burns et al., 2022; Mallen et al., 2023); our $\mathrm{Recov}(I; Z)$ metric complements this

line by measuring supervised task-specific decodability at a strict pre-emission boundary, alongside work on faithfulness gaps between internal computation and external rationales (Lanham et al., 2023; Turpin et al., 2023; Xia et al., 2025). Second, we relate to **pre-generation estimation and uncertainty**: semantic entropy and CLOTHO estimate uncertainty or adequacy from sampling and hidden states (Kuhn et al., 2023; Yoon et al., 2025), whereas $H_{\text{int}}(I)$ and $\text{Recov}(I; Z)$ provide single-pass diagnostics before linguistic commitment, unlike output-side calibration methods (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023). Our response-format findings are also consistent with evidence that strict output constraints can impair reasoning performance (Tam et al., 2024), especially in multiple-choice settings where our compliance-conditioned and restricted-choice results suggest content-versus-symbol binding failures. Third, many **reasoning-enhancing methods** can be viewed as interventions on the pre-emission state, including CoT prompting (Wei et al., 2022), rationale bootstrapping (Zelikman et al., 2022, 2024), process supervision (Lightman et al., 2023; Uesato et al., 2022), self-refinement (Madaan et al., 2023), and test-time adaptation (Sun et al., 2019; Gandelsman et al., 2022).

5 Limitations

Our results should be interpreted within a bounded empirical scope: we evaluate three instruction-tuned model families in the 7–8B range, three text-only reasoning benchmarks, and a fixed greedy decoding policy. This supports controlled comparison, but does not address larger-scale models, alternative architectures, multimodal settings, or sampling-based collapse policies; because greedy decoding isolates prompt-conditioning effects at the commitment boundary but departs from common CoT practice, our reported CoT effects may also be conservative relative to typical deployment settings. Likewise, $n = 200$ items per model–benchmark cell is sufficient to reveal robust qualitative patterns, but not small effects or higher-order interactions. The proposed metrics also have important scope conditions: intention entropy $H_{\text{int}}(I)$ is computed over the full next-token vocabulary and is best interpreted as local pre-emission generative dispersion rather than calibrated decision uncertainty, especially in multiple-choice settings; restricted-choice analogues are therefore reported in Appendix B.6. Recoverability $\text{Recov}(I; Z)$ measures linear accessibility of task-relevant signal, not competence or guaranteed correct expression, and linear decodability alone does not establish causal use during collapse. Our findings are therefore correlational rather than causal. Finally, Babble is only a partial verbosity control because it changes both length and semantics, and activation extraction uses 4-bit quantization for practical efficiency; a cross-precision robustness check for intention entropy is reported in Appendix B.5, while broader cross-precision and compute-matched analyses are left to future work.

6 Conclusion

We introduced *Intention Collapse* as a practical framework for studying the boundary between pre-emission internal state and final output in language models. Using three lightweight diagnostics—intention entropy, effective dimensionality, and recoverability—we evaluated three model families across three benchmarks and three prompting regimes.

Our results show that CoT is not always beneficial: it helps on GSM8K, hurts on ARC-Challenge, and has mixed effects on AQUA-RAT. We also find that these effects vary across model families, and that useful pre-emission signal can sometimes remain present even when final accuracy declines.

These findings suggest that understanding LLM reasoning requires analyzing not only outputs, but also how pre-emission representations are converted into final decisions. In this sense, Intention Collapse is a lightweight empirical framework for studying when internal signal is present, how prompting reshapes it, and when that signal fails to survive collapse.

References

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

Paul Christiano and Mark Xu. 2021. ARC's first technical report: Eliciting latent knowledge. Technical report, Alignment Research Center.

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

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Neural Information Processing Systems (NeurIPS)*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A. Efros. 2022. Test-time training with masked autoencoders. In *Advances in Neural Information Processing Systems*. NeurIPS 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 81 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and 1 others. 2023. Mistral 7b. *arXiv preprint*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

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

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

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

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. 2023. Eliciting latent knowledge from quirky language models. *arXiv preprint*.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. 2019. Test-time training with self-supervision for generalization under distribution shifts. *arXiv preprint arXiv:1909.13231*.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *arXiv preprint arXiv:2408.02442*.

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

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

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Tianle Xia, Liang Ding, Guojia Wan, Yibing Zhan, Bo Du, and Dacheng Tao. 2025. Improving complex reasoning over knowledge graph with logic-aware curriculum tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12881–12889.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Juyeon Yoon, Somin Kim, Robert Feldt, and Shin Yoo. 2025. Clotho: Pre-generation adequacy estimation for language models. *arXiv preprint arXiv:2509.17314*.

Eric Zelikman, Grady R. Harik, Yuhuai Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. 2024. Quiet-STaR: Language models can teach themselves to think before speaking. *arXiv preprint*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488.

A Appendices

A.1 Code Availability

Code will be released upon publication.

The repository includes: (i) scripts for activation extraction at the pre-collapse boundary (Section 2), (ii) metric computation for $H_{\text{int}}(I)$ and $\dim_{\text{eff}}(I)$, (iii) probe training and evaluation for $\text{Recov}(I; Z)$, (iv) parsers and compliance diagnostics for multiple-choice benchmarks, and (v) notebooks to reproduce figures and consolidated tables for the 3×3 matrix reported in Section 3.

We will release code and prompt templates sufficient to reproduce all reported tables and figures upon publication.

A.2 Models and Benchmarks

We evaluate three instruction-tuned open-weight models in the 7–8B range:

- Mistral-7B-Instruct
- Llama-3.1-8B-Instruct
- Qwen-2.5-7B-Instruct

All runs use greedy decoding (temperature 0.0) and practical inference settings (4-bit quantization for activation extraction). For each model–benchmark cell we evaluate $n = 200$ randomly sampled items.

Benchmarks:

- **GSM8K (free-response math)**: numeric final answers.
- **ARC-Challenge (multiple-choice)**: one option among a fixed set.
- **AQUA-RAT (multiple-choice math)**: one option among a fixed set.

A.3 Prompt Templates

We used separate prompt templates for free-response (GSM8K) and multiple-choice (ARC/AQUA) to ensure that the commitment format is explicit. All prompts are applied within the same three regimes: Baseline, CoT, and Babble.

GSM8K: Baseline (Direct Answer).

```
Solve this math problem. Give only the
final numerical answer.

Problem: question Answer:
```

GSM8K: CoT (Chain-of-Thought).

```
Solve this math problem step by step.
Show your reasoning, then give the final
answer after####.

Problem: question Solution:
```

GSM8K: Babble (Length Control).

```
Given this math problem, write a long
stream of consciousness about numbers,
calculations, and mathematical concepts.
Do NOT solve the problem - just write
loosely related mathematical musings for
about 100 words.

Problem: question Stream of
consciousness:
```

Multiple-choice: Baseline (Direct Answer).

```
Choose the correct option. Reply with
only one letter from the available
choices.

Question: question Choices: choices
Answer:
```

Multiple-choice: CoT (Chain-of-Thought).

```
Think step by step, then choose the
correct option. At the end, reply
with only one letter from the available
choices.

Question: question Choices: choices
Reasoning: Final answer:
```

Multiple-choice: Babble (Length Control).

```
Write a long stream of consciousness
about the topic and the possible answers.
Do NOT solve the problem. At the end,
still output exactly one letter from the
available choices.

Question: question Choices: choices
Stream of consciousness: Final answer:
```

A.4 Answer Extraction, Parsing, and Compliance

For GSM8K, we extract the final numeric answer using standard normalization rules (e.g., last occurring numeric token in the final answer span after the delimiter). For multiple-choice benchmarks, we extract the final option letter using a strict parser (single valid option among the allowed set). If parsing fails (no valid single option), the output is marked incorrect. To diagnose format effects, we report compliance rates for multiple-choice tasks (percentage of generations containing exactly one valid option token at the end) alongside accuracy. This is particularly important in settings where CoT reduces accuracy, since some failures may reflect commitment/compliance breakdown rather than lack of latent signal or lack of recoverable internal information.

A.5 Decoding Parameters

All three conditions used greedy decoding to eliminate stochasticity:

- **Temperature**: 0.0 (greedy decoding)
- **Max tokens (Baseline)**: 50
- **Max tokens (CoT/Babble)**: 512

Greedy decoding ensures that differences in $H_{\text{int}}(I)$, $d_{\text{eff}}(I)$, and $\text{Recov}(I; Z)$ are attributable as cleanly as possible to prompt conditioning and the resulting generation regime rather than to stochastic sampling variation at the commitment boundary. At the same time, the larger budgets for CoT/Babble introduce a compute confound; we therefore interpret length controls as partial, and we treat compute-matching as a priority for future work.

A.6 Pre-collapse Extraction and Probe Protocol (Summary)

We extract I at the last prompt position immediately before the first output token is selected, identically across Baseline, CoT, and Babble. We compute: (i) $H_{\text{int}}(I)$ from the next-token distribution at that position; (ii) $d_{\text{eff}}(I)$ via per-layer PCA participation ratios with variance-weighted aggregation; activations are mean-centered per layer before PCA (no z-scoring across features, as variance differences across dimensions are preserved for the participation ratio calculation); and (iii) $\text{Recov}(I; Z)$ via ℓ_2-regularized logistic regression probes with item-level splits (60/20/20) and bootstrap confidence intervals on AUROC. Item-level splits are constructed once per benchmark using item IDs, and the resulting train/validation/test partitions are then shared across all three regimes. As a consequence, an item assigned to the test set under one regime is also assigned to the test set under the other regimes, eliminating cross-regime item leakage by construction. This design isolates representational differences across regimes from item-difficulty variation, since the same items appear in each split under all conditions. Additional robustness checks for d_{eff} (subsampling and intrinsic-dimension cross-checks) and for probes (train–test gaps, label-shuffle baselines, and cross-regime transfer) are reported in Appendix B.

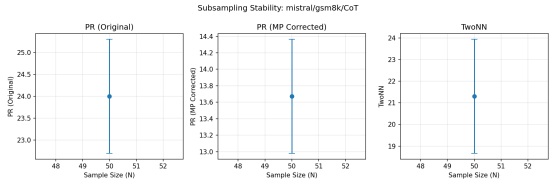

Figure 4: Subsampling stability of $d_{\text{eff}}^{\text{agg}}$ across item counts N′ for the Mistral/GSM8K/CoT cell (selected as the most conservative test given its large entropy shift). Error bars indicate standard deviation across R = 20 repeated subsamples. The narrow absolute range of dagg eff values across N′ confirms that the participation ratio estimate is stable at N = 200

B Robustness and Additional Analyses

B.1 Effective dimensionality robustness

Subsampling stability. To assess finite-sample sensitivity in the $d \gg N$ regime, we recompute d_{eff} under repeated subsampling of items within each model–benchmark–regime cell. Specifically, for each $N' \in \{25, 50, 100, 150, 200\}$,we sample N' items without replacement (repeated R = 20 times) and report the mean and standard deviation of $d_{\text{eff}}^{\text{agg}}$. Figure 4 shows results for the Mistral/GSM8K/CoT cell as a representative example; this cell was selected because it exhibits the largest cross-regime deff shift in the matrix, making it the most conservative test of subsampling stability. Figure 4 shows that the qualitative cross-regime trends emphasized in the main text are stable across subsample sizes.

Cross-checks with intrinsic-dimension estimators. Participation ratio is spectrum-based and can be biased by finite-sample covariance estimation. As complementary checks, we com-

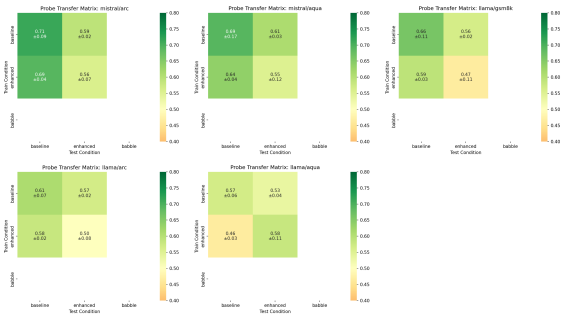

Figure 5: Cross-regime transfer matrices (AUROC). Systematic degradation under transfer indicates representational shift across regimes even when within-regime recoverability is above chance.

pare trends against intrinsic-dimension estimators derived from local neighborhood structure (TwoNN) and maximum-likelihood estimators (Levina–Bickel). While these estimators differ in absolute scale and assumptions, they corroborate the direction of regime shifts in the conditions where d_{eff} changes most strongly.

B.2 Probe robustness checks

We evaluate probe robustness in three ways: (i) train–test AUROC gaps under the selected regularization strength C, (ii) a label-shuffle sanity check (training labels permuted; expected AUROC ≈ 0.5), and (iii) reporting a fixed pre-specified layer set to limit researcher degrees of freedom.

Note on bootstrap CIs under severe imbalance. For highly imbalanced cells (e.g., AQUA baseline with very few positives), naive bootstrap resamples may contain only one class, making AUROC undefined and yielding NaN confidence intervals. In such cases we omit the CI or report it as unavailable; this does not affect the point estimate and is consistent with standard AUROC practice under extreme imbalance.

B.3 Cross-regime probe transfer

To test whether recoverable information corresponds to shared predictive axes or regime-specific encodings, we evaluate cross-regime transfer: train on one regime and test on another while preserving item splits. Figure 5 summarizes transfer matrices (AUROC) for the available model/benchmark settings.

B.4 Compact cross-cell summary results

For transparency, Table 3 reports a compact cross-cell summary used in the main discussion: Baseline and CoT accuracy, Baseline and CoT intention entropy, the resulting entropy-shift pattern, and CoT probe AUROC with 95% bootstrap confidence intervals.

B.5 Quantization robustness

Activation extraction uses 4-bit quantization for practical efficiency, which could in principle distort the geometry of hidden states relative to full-precision computation. To bound this risk, we conducted an independent replication of the Mistral-7B / GSM8K cell on different hardware using full-precision weights (BF16). For intention entropy H_{int}, accuracy differed by 2.5 percentage points (11.0% vs. 13.5% baseline; 50.0% vs. 52.5% CoT), and H_{int} differed by less than 0.08 bits across all conditions; critically, the rank ordering of entropy across regimes (baseline $>$ babble $\approx$ CoT) was preserved exactly. We did not retain the activation tensors from this replication and therefore cannot report cross-precision comparisons for d_{eff} or probe AUROC on the same cell; a complete cross-precision comparison of all three diagnostics across the 3×3 matrix remains future work. We note, however, that participation-ratio dimensionality is invariant to global scale and is dominated by the dominant eigenvalues, both of which should be relatively robust to 4-bit weight quantization; and that linear probes are trained over the same activations they evaluate, so any systematic distortion would affect train and test alike. We flag these as informed expectations, not measured results, and treat the lack of full cross-precision evidence for d_{eff} and Recov as a scope condition of the present study.

B.6 Restricted-choice diagnostics for multiple-choice settings

A natural concern with full-vocabulary uncertainty metrics is that, in multiple-choice settings, admissible answers occupy only a very small subset of the output space. To address this concern, we computed supplementary *restricted-choice* diagnostics for the multiple-choice benchmarks. Starting from the pre-emission hidden state, we applied the model's final normalization and LM head, retained logits only for the valid answer options for the corresponding benchmark, renormalized over that subset, and derived restricted-choice metrics from the resulting distribution.

These diagnostics are intended as a robustness check rather than a replacement for the main anal-

Table 2: Probe robustness analysis. "baseline" denotes the direct-answer regime and "enhanced" denotes the CoT regime used throughout the main text. We report train and test AUROC with 95% bootstrap confidence intervals, label-shuffle sanity checks (expected ≈ 0.5), and train–test gap (positive values indicate potential overfitting).

Model	Benchmark	Condition	Train	Test [95% CI]	Shuffle	Gap
Mistral	GSM8K	baseline	1.00	0.57 [0.37, 0.79]	0.53	+0.43
Mistral	GSM8K	enhanced	0.98	0.51 [0.35, 0.66]	0.51	+0.47
Mistral	ARC	baseline	1.00	0.86 [0.75, 0.95]	0.53	+0.14
Mistral	ARC	enhanced	1.00	0.48 [0.32, 0.63]	0.54	+0.52
Mistral	AQUA	baseline	1.00	0.74 [–, –]	0.57	+0.26
Mistral	AQUA	enhanced	1.00	0.75 [0.58, 0.89]	0.57	+0.25
Llama	GSM8K	baseline	1.00	0.73 [0.53, 0.91]	0.49	+0.27
Llama	GSM8K	enhanced	1.00	0.50 [0.32, 0.68]	0.60	+0.50
Llama	ARC	baseline	0.95	0.56 [0.42, 0.71]	0.52	+0.39
Llama	ARC	enhanced	1.00	0.50 [0.36, 0.66]	0.49	+0.50
Llama	AQUA	baseline	0.97	0.63 [0.46, 0.79]	0.53	+0.33
Llama	AQUA	enhanced	0.98	0.46 [0.30, 0.64]	0.57	+0.52
Qwen	GSM8K	baseline	1.00	0.75 [0.57, 0.88]	0.55	+0.25
Qwen	GSM8K	enhanced	0.97	0.60 [0.45, 0.75]	0.49	+0.37
Qwen	ARC	baseline	0.95	0.82 [0.71, 0.91]	0.56	+0.13
Qwen	ARC	enhanced	1.00	0.71 [0.56, 0.84]	0.51	+0.29
Qwen	AQUA	baseline	1.00	0.58 [0.36, 0.80]	0.52	+0.42
Qwen	AQUA	enhanced	0.95	0.62 [0.46, 0.75]	0.52	+0.34

Note: For severely imbalanced cells, naive bootstrap resamples may contain only one class, making AUROC undefined; we report CI as unavailable in those cases.

ysis. In particular, they test whether some of the uncertainty patterns identified with full-vocabulary metrics remain visible when attention is restricted to admissible answer options only. Table 4 reports a compact subset of these diagnostics: final parsed accuracy, restricted-choice argmax accuracy, their difference (*Gap*), and agreement between the restricted-choice argmax and the extracted final answer.

The main pattern is asymmetric across regimes. In all baseline cells with full logging, the restricted-choice argmax accuracy is at least as high as the final parsed accuracy, yielding a non-negative hidden-correct / expressed-wrong gap. This indicates that, in those cells, task-relevant answer information is sometimes more recoverable from the pre-emission state over admissible options than from the final parsed output. Under CoT, however, the pattern becomes mixed: some cells retain a positive gap, whereas others show negative gaps together with lower agreement between restricted-choice argmax and extracted answers. This instability is itself informative: it suggests that CoT prompting can alter the first-token distribution in ways that are partly orthogonal to answer quality, consistent with the format-dependence reported in Section 3.

Restricted-choice entropy provides additional evidence that pre-emission states carry decision-relevant signal beyond what full-vocabulary entropy captures. In 10 of the 12 cells with full logging, H_{restr} is higher on incorrectly answered items than on correctly answered ones (mean discrimination $+0.144$ bits), indicating that the model's pre-emission distribution over admissible options is generally more dispersed when it is about to fail. The two exceptions (Mistral/AQUA CoT and Qwen/AQUA CoT) coincide with cells in which the restricted-choice argmax becomes dominated by a single option, consistent with the first-token bias pattern noted above. These results suggest that, in multiple-choice settings, task-restricted entropy can serve as a more decision-aligned diagnostic than full-vocabulary entropy, while still inheriting

Table 3: Main results across models and benchmarks. We report per-cell accuracy with Wilson 95% confidence intervals, pre-collapse entropy $H_{\text{int}}(I)$ in bits, and probe AUROC with 95% CI for the CoT condition. Arrows indicate change from Baseline to CoT.

Model	Benchmark	Acc (B→CoT)	H_{int} (B→CoT)	Pattern	AUROC [95% CI]
Mistral-7B	GSM8K	11.0 [7.4, 16.1] → 50.0 [43.1, 56.9]	2.39→0.37	CF	0.56 [0.49, 0.64]
	ARC-Challenge	68.0 [61.2, 74.1] → 66.5 [59.7, 72.7]	0.94→0.69	CF	0.52 [0.43, 0.60]
	AQUA-RAT	6.5 [3.8, 10.8] → 13.0 [9.0, 18.4]	2.85→0.49	CF	0.54 [0.42, 0.66]
LLaMA-3.1-8B	GSM8K	19.0 [14.2, 25.0] → 79.5 [73.4, 84.5]	1.57→4.53	EC	0.56 [0.46, 0.65]
	ARC-Challenge	55.0 [48.1, 61.7] → 45.0 [38.3, 51.9]	0.98→4.27	EC	0.53 [0.45, 0.61]
	AQUA-RAT	29.0 [23.2, 35.6] → 14.5 [10.3, 20.0]	3.27→4.56	EC	0.61* [0.51, 0.72]
Qwen-2.5-7B	GSM8K	15.5 [11.1, 21.2] → 41.5 [34.9, 48.4]	0.67→2.21	EC	0.65* [0.57, 0.73]
	ARC-Challenge	81.0 [75.0, 85.8] → 52.0 [45.1, 58.8]	0.67→1.57	EC	0.69* [0.61, 0.76]
	AQUA-RAT	23.0 [17.7, 29.3] → 63.5 [56.6, 69.9]	1.24→2.53	EC	0.58 [0.49, 0.66]

CF = Collapse-first (entropy decreases); EC = Explore-then-commit (entropy increases). Accuracy intervals are Wilson 95% confidence intervals computed per cell ($n = 200$). * indicates AUROC confidence intervals excluding chance (0.5). AUROC values in this table were recomputed using a refined probe protocol; full robustness details including train/test gaps and label-shuffle checks are reported in Table 2 (Appendix B.3).

Table 4: Restricted-choice diagnostics for multiple-choice settings. H_{full} = full-vocabulary intention entropy (bits); H_{restr} = entropy restricted to valid option tokens; H_{disc} = mean H_{restr} on incorrect minus correct items (positive values indicate that restricted entropy discriminates correctness); RC argmax = restricted-choice argmax accuracy; Gap = RC argmax minus parsed; RC-agr = agreement between restricted-choice argmax and extracted final answer.

Model	Bench.	Regime	Parsed	H_{full}	H_{restr}	H_{disc}	RC argmax	Gap	RC-agr
Mistral	AQUA	Base	0.065	2.851	1.813	+0.124	0.235	+0.170	0.346
Mistral	AQUA	CoT	0.130	0.488	0.655	-0.045	0.200	+0.070	0.333
Mistral	ARC	Base	0.680	0.937	0.427	+0.448	—	—	0.975
Mistral	ARC	CoT	0.665	0.690	0.535	+0.026	—	—	0.500
LLaMA-3.1-8B	AQUA	Base	0.290	3.275	2.029	+0.070	0.300	+0.010	0.898
LLaMA-3.1-8B	AQUA	CoT	0.145	4.563	1.152	+0.202	0.195	+0.050	0.396
LLaMA-3.1-8B	ARC	Base	0.550	0.978	0.816	+0.192	0.745	+0.195	0.740
LLaMA-3.1-8B	ARC	CoT	0.450	4.273	1.169	+0.097	0.340	-0.110	0.508
Qwen-2.5-7B	AQUA	Base	0.230	1.236	0.830	+0.193	0.385	+0.155	0.699
Qwen-2.5-7B	AQUA	CoT	0.635	2.527	0.878	-0.016	0.195	-0.440	0.200
Qwen-2.5-7B	ARC	Base	0.810	N/A	0.091	+0.155	0.875	+0.065	0.920
Qwen-2.5-7B	ARC	CoT	0.520	1.568	0.266	+0.140	0.245	-0.275	0.600

RC argmax was not retained for the earlier Mistral ARC runs, for which only parsed accuracy and argmax–extraction agreement are available.

some of the same regime-dependent artifacts that affect restricted-choice argmax.

For this reason, we do not use restricted-choice diagnostics as a primary evaluation axis and do not base the paper's main claims on them. Their value is narrower but still useful: they show that the full-vocabulary results are not merely artifacts of scoring over irrelevant tokens, while also revealing that choice-restricted diagnostics can become regime-sensitive under some CoT prompting conditions. We therefore report H_{restr} as a supplementary diagnostic rather than as a replacement for the main metrics.