Is It Necessary and Robust? Decomposing LLM's Reasoning Quality

Anonymous ACL submission

Abstract

Scaling up language models does not always improve the quality of their reasoning. We propose a framework that decomposes reasoning quality into *necessity*, which captures whether a reasoning trace changes the model's answer distribution, and *robustness*, which captures whether that distribution remains stable under controlled perturbations. Across the model families and benchmarks we examine, explicit reasoning is most necessary at intermediate model scales, while robustness to reasoning perturbations is weakest at intermediate scales. The framework accommodates commonly used linguistic attributes of reasoning traces such as completeness and redundancy. Code is available at this url.

1 Introduction

Modern large language models (LLMs) often generate multi-step reasoning to produce an answer. As these traces become central to evaluating LLM performance, an important question is whether the reasoning itself is of high quality. However, a general framework for quantifying reasoning quality across models and tasks remains underdeveloped.

In this paper, we propose a task- and model-agnostic framework for measuring LLM reasoning quality. Our framework is grounded in an inferentialist view of reasoning, according to which a reason is meaningful insofar as it plays a consequential inferential role and remains stable under relevant counterfactual variation (Brandom, 2000, 2008). We thus decompose reasoning quality into two components: **necessity** and **robustness**. Necessity asks whether the reasoning trace materially changes the model's answer distribution relative to the question alone. Robustness asks whether the answer distribution remains stable when the trace is perturbed. Necessity captures whether a reasoning trace makes the model's answer distribution closer to a reference answer distribution, measured as the reduction in Kullback–Leibler divergence from the question-only distribution to the trace-conditioned distribution. We also formalize robustness through interventions on the reasoning trace and show that event-level reasoning robustness is lower bounded by a Jensen–Shannon penalty between the original and perturbed answer distributions. Together, we provide a task-agnostic method to quantify whether reasoning is consequential and stable.

The framework yields a qualitative prediction that reasoning quality need not improve monotonically with model scale. *Necessity may be inverse U-shaped*. Small models may fail to use the reasoning trace effectively, while sufficiently strong models may solve the task largely from the question alone, reducing the marginal value of explicit reasoning. *Robustness may instead be U-shaped*. Weak models can appear stable because they do not meaningfully use the trace; intermediate models may become more fragile as they begin to rely on explicit reasoning; and large models may gain stability by recovering from local perturbations. Thus, scaling may change the *composition* of reasoning quality rather than uniformly improving it.

We test these predictions across extensive model classes, sizes, and tasks. Necessity and robustness vary substantially across settings. The results also confirm the non-monotonic relationships between model scale and observed reasoning quality components. Additional theory and experiments show that linguistic properties of reasoning traces, such as completeness and redundancy, can be understood under the same necessity–robustness framework.

Our contributions are as follows. First, we introduce a task- and model-agnostic framework that decomposes reasoning quality into necessity and robustness and propose simple empirical methods to quantify each. Second, we show that reasoning quality varies non-monotonically with model scale, indicating that scaling up does not uniformly improve reasoning. Third, we show that our frame-

work can accommodate commonly used linguistic attributes such as completeness and redundancy.

2 Preliminaries

Let X denote a question, S a reasoning trace, and $A \in \mathcal{A}$ the model's answer. For a model m, let $P_m(A \mid X)$ and $P_m(A \mid X, S)$ denote the answer distributions without and with the reasoning trace.

2.1 Necessity as Information Gain

Necessity is whether the reasoning trace materially changes the model's answer distribution relative to the question alone. We formulate necessity as the conditional mutual information between the answer and the reasoning trace given the question.

Definition 1. For model m, define $P_m^0 = P_m(A \mid X)$ and $P_m^S = P_m(A \mid X, S)$. Then, the **necessity** of reasoning trace S is

$$\mathrm{KL}\big(P_m(A) \,\|\, P_m^0\big) - \mathrm{KL}\big(P_m(A) \,\|\, P_m^S\big).$$

This suggests that a reasoning trace is *necessary* when conditioning on it induces a substantial belief update over answers. By setting $A = E^\star$, where $E^\star$ is the event that the answer is correct, we also define **correctness-adjusted necessity**.

2.2 Robustness as Intervention Immunity

We define robustness as whether a model's answer remains stable when its reasoning trace is perturbed in controlled ways. This is because a high-quality reasoning trace should have even information density and be immune to a random perturbation.

Let $\{\mathcal{I}_\lambda\}_{\lambda \in [0,1]}$ be a family of intervention operators, where larger λ corresponds to stronger perturbation. For a fixed trace S, let $S' \sim \mathcal{I}_\lambda(S)$ denote a perturbed version of the trace. For any event $A \subseteq \mathcal{A}$ (e.g., the model's answer is correct), define the event-level *instability* of reasoning $\Delta_{m,\lambda}(X, S; A)$ by

$$\mathbb{E}_{S' \sim \mathcal{I}_\lambda(S)} \left[\left| P_m(A \mid X, S) - P_m(A \mid X, S') \right| \right].$$

Robustness is the *opposite* notion.

Definition 2. For model m, question X, trace S, and perturbation strength λ, $\mathcal{P}_{m,\lambda}(X, S)$ is defined as follows (JS is Jensen-Shannon divergence).

$$\mathbb{E}_{S' \sim \mathcal{I}_\lambda(S)} \left[\mathrm{JS}\big(P_m(A \mid X, S),\ P_m(A \mid X, S')\big) \right].$$

We define **robustness** as $-1 \times \mathcal{P}_{m,\lambda}(X, S)$, measured by trace perturbations. We identify the upper bound of $\mathcal{P}_{m,\lambda}(X, S)$ as follows.

Proposition 1. *For any event $E \subseteq \mathcal{A}$, define $p := P_m(A \mid X, S)$ and $q := P_m(A \mid X, S')$. Then,*

$$\begin{aligned} \Delta_{m,\lambda}(X, S; E) &\le \mathbb{E}_{S' \sim \mathcal{I}_\lambda(S)} \left[\sqrt{2\,\mathrm{JS}(p, q)} \right] \\ &\le \sqrt{2\,\mathcal{P}_{m,\lambda}(X, S)}. \end{aligned}$$

Proof. The result follows by applying Pinsker inequality to the JS penalty. See Appendix G.1. □

2.3 RQ is Non-Monotonic to Model Scale

Let $m \in \mathcal{M}$ index model capability. We summarize the effect of scale on necessity by $h_N(m) = \ell(m)\big(1 - \sigma(m)\big)$, where $\ell(m)$ captures the model's ability to use the trace and $\sigma(m)$ captures question self-sufficiency. Similarly, we summarize the effect of scale on instability (i.e., opposite of robustness) by $h_I(m) = r(m)\big(1 - u(m)\big)$, where $r(m)$ captures reliance on explicit intermediate reasoning and $u(m)$, recoverability under perturbation. Increases in $\ell(m)$ and $r(m)$ make reasoning more consequential, while increases in $\sigma(m)$ and $u(m)$ reduce the marginal value of explicit reasoning.

Theorem 1. *Fix (X, S, λ). Suppose trace usage $\ell(m)$, reliance $r(m)$, question self-sufficiency $\sigma(m)$, and recoverability $u(m)$ all increase with model capability. If the relative growth of $\ell(m)$ single-crosses that of $\sigma(m)$, and the KL-based necessity measure is order-preserving in the trace-induced update magnitude, then necessity is single-peaked in m. If the relative growth of $r(m)$ single-crosses that of $u(m)$, then $\mathcal{P}_{m,\lambda}$(X,S) is also single-peaked in m.*

Proof. For empirical demonstration of single-crossing property, see Appendix C.2. For formal proof, see Appendix G.2. □

Theorem 1 predicts that necessity and model scale have an inverse-U-shaped relationship, while robustness and scale have a U-shaped relationship.

3 Experiments

3.1 Data and Measurement

Models and datasets. We evaluate on three benchmark datasets: Google-Proof QA (GPQA) (Rein et al., 2023), MMLU (Hendrycks et al., 2021a), and MATH (Hendrycks et al., 2021b). See Appendix A.1 for dataset descriptions. We evaluate across 10 models from 2 model families: `Qwen 2.5` (0.5B, 1.5B, 3B, 7B, 14B and 32B) (Yang et al., 2025b) and `Gemma 3` (Gemma Team et al., 2025)

(1B, 4B, 12B, and 27B). This allows *within-family comparisons* across model scales. In the main experiment, we use instruction-tuned models since reasoning-tuned models often make the question-only distribution uninterpretable. See Appendix D.4 for detailed explanations and an ablation using a reasoning model, Qwen 3 (Yang et al., 2025a).

Measuring necessity and robustness. We empirically measure correctness-adjusted necessity and robustness as follows. Let δ_y denote the one-hot distribution on the correct answer and $P_m^{S'} = P_m(A \mid X, S')$. We define

$$\text{Necessity}_m := \text{KL}(\delta_y \| P_m^0) - \text{KL}(\delta_y \| P_m^S) \quad (1)$$

$$\text{Robustness}_{m,\lambda} := -\text{JS}(P_m^S, P_m^{S'}). \quad (2)$$

To measure robustness, we construct a perturbed trace S' by independently selecting a fraction $\lambda = 10\%$ of trace tokens and masking each selected token with `<redacted_UNK>`. We use the upper bound identified in Proposition 1 as an empirical measure of robustness. Appendix A.2 describes the details used in the measurement. Appendix A.3 validates that the JS upper bound is empirically tightest for correct events.

Obtaining answer token distribution is straightforward for multiple-choice questions (e.g., MMLU and GPQA). For open-ended tasks (e.g., MATH), we approximate answer distributions by teacher forcing (Bengio et al., 2015; Mihaylova and Martins, 2019). Specifically, we evaluate the model's token-level probability assigned to the ground-truth answer under two contexts: question-only and question-plus-reasoning. We tokenize the ground-truth answer, measure the conditional probability at each token position, and average the negative log-likelihoods to obtain a single KL score per question (see Appendix A.4). We then aggregate the scores to subtask (e.g., chemistry, physics in MMLU) or benchmark level in downstream tasks.

Necessity and robustness are distinct. Appendix Table 4 shows that necessity and robustness vary substantially across benchmarks and models. Appendix B.1 shows empirical distributions of the two measures. The two metrics do not have statistically significant correlation and they represent distinct dimensions of reasoning quality. Traces can be consequential but fragile, stable but unnecessary, both, or neither.

3.2 Model Scale and Reasoning Quality

Necessity is inverse-U-shaped in model scale. Figure 1 shows the relationship between model scale and correctness-directed necessity aggregated at the model-benchmark level. We min-max scale the metric for interpretability. Across models and benchmarks, we find an inverse U-shaped relationship between necessity and scale. We fit the following mixed-effects quadratic model at the subtask level: $\text{Metric} = \beta_1 \log(\text{Scale}) + \beta_2 \log(\text{Scale})^2 + (1|\text{Subtask}) + (1|\text{Benchmark})$. Both models show a significant negative β_2. Appendix B.2 elaborates the statistical details and introduces more statistical tests to validate the non-monotonic relationship.

This pattern is consistent with Theorem 1. Small models do not use the reasoning trace effectively, so conditioning on the trace produces only limited improvement. Intermediate models use the trace more effectively, making explicit reasoning most consequential. Larger models become question-self-sufficient, reducing the value of explicit trace.

Robustness is U-shaped in model scale. Similarly, in Figure 1, we find a U-shaped relationship between robustness and model scale. Statistically, we find a positive β_2 in the mixed-effects regression and subtask bootstraps. The smallest models appear relatively stable because they do not meaningfully use the reasoning trace. Intermediate models are least robust because they rely on explicit reasoning but have not yet developed sufficient recoverability under perturbation. The largest models regain robustness, consistent with the idea that these models can recover from local perturbations.

Additional results. Appendix D.1 changes the masking ratio from 10% to 5%, 20%, and 30%. A higher ratio increases the JS penalty (i.e., lower robustness), yet the U-shaped relationship between robustness and model scale remains robust. Appendix D.2 reports the necessity and robustness measures at the model-subtask level. We consistently find a non-monotonic relationship between reasoning quality components and model scale at the subtask level. Appendix D.3 confines the sample to when necessity has a positive value (i.e., distribution being closer to the correct answer). The non-monotonic relationships persist across all specifications. Appendix D.4 shows that the results also hold with a reasoning-tuned model, Qwen 3.

Jointly, the results suggest that scaling up a language model does not always lead to an improve-

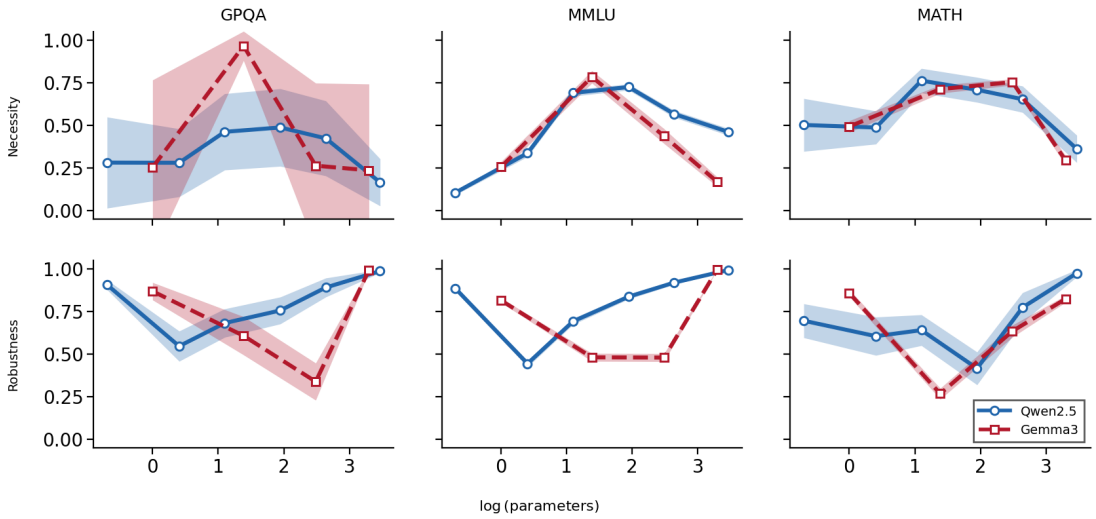

Figure 1: **Necessity and robustness aggregated at the model-benchmark level**. We average question-level min–max normalized necessity at the model-benchmark level (top) and robustness (bottom) vs. log model size. Shaded bands are 99% confidence intervals. The underlying data of this figure is in Appendix Table 4.

ment in reasoning quality. Instead, *scale changes the composition of reasoning quality* – larger models rely less on traces reducing the need for explicit reasoning, while they become more robust to perturbations when such traces are used.

3.3 Other Linguistic Attributes

Our framework hosts other common reasoning quality measures like completeness or redundancy.

Completeness. A complete reasoning trace is one for which admissible perturbations do not change the induced answer distribution (Qiu et al., 2025). Appendix G.3 proves that, if a trace is complete, both event-level and the JS-based penalty are zero. Appendix E.1 validates this by showing that deleting reasoning steps increases the JS penalty.

Redundancy. Redundancy means that the trace contains repeated reasoning steps that are conceptually unnecessary to render the conclusion (Prasad et al., 2023). In Appendix G.4, we show that redundancy increases the upper bound of Proposition 1. Each repeated location can worsen robustness, and in the worst case these contributions accumulate rather than cancel. In Appendix E.2, we find consistent empirical evidence.

4 Related Work

There has been growing interest in improving LLM reasoning through intermediate traces, including scratchpads, chain-of-thought prompting, zero-shot CoT, self-consistency, and rationale-based bootstrapping (Kaplan et al., 2020; Wei et al., 2022b; Kojima et al., 2022; Wei et al., 2022a; Wang et al., 2023; Zelikman et al., 2022). Studies have shown that explicit reasoning tends to improve performance, particularly in mathematical, symbolic, or deliberation-friendly settings (Sprague et al., 2025; Liu et al., 2025). Beyond accuracy gain, a growing literature evaluates the quality of reasoning traces themselves, using intervention-based, counterfactual, causal, distributional, benchmark-based, step-level reasoning rigidity-based, and survey-based approaches (Turpin et al., 2023; Lanham et al., 2023; Madsen et al., 2024; Siegel et al., 2024; Paul et al., 2024; Matton et al., 2025; Tutek et al., 2025; Lee and Hockenmaier, 2025; Basu and Chakraborty, 2026). Instead, we provide a *task- and model-agnostic formal framework that decomposes reasoning quality into necessity and robustness* and show how both components vary with model scale.

5 Conclusion

We decompose LLM reasoning quality into necessity and robustness. Theory links these components to answer-distribution quantities and identifies conditions under which scaling need not improve reasoning monotonically. Empirically, our results support this prediction. These findings suggest that reasoning quality is not a single monotone property of scale, but a composition of distinct factors.

Limitations

Our experiments focus on open-weight model families because our framework requires repeated access to token-level answer distributions under question-only, trace-conditioned, and perturbed-trace conditions. Many proprietary systems restrict access to full log-probabilities, which makes systematic evaluation of closed-source models difficult. In addition, although we evaluate multiple model families and reasoning benchmarks, the results should not be interpreted as universal laws of scaling. The shape of necessity and robustness may depend on training data, post-training procedures, prompt format, and benchmark structure. Lastly, our Theorem relies on mild empirical assumptions. Although validated through our experiments, a violation would invalidate the theorem.

AI Usage Disclosure

AI tools were used for grammar editing and code efficiency checks. The authors are responsible for all substantive ideas, analyses, interpretations, and conclusions.

References

Abhinaba Basu and Pavan Chakraborty. 2026. Measuring and curing reasoning rigidity: from decorative chain-of-thought to genuine faithfulness. *arXiv preprint arXiv:2603.22816*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28.

Robert B. Brandom. 2000. *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press.

Robert B. Brandom. 2008. *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford University Press.

Gemma Team, Aishwarya Kamath, Johan Ferret, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35.

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*.

Jinu Lee and Julia Hockenmaier. 2025. Evaluating step-by-step reasoning traces: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1789–1814.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2025. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 38489–38517.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337.

Katie Matton, Robert Osazuwa Ness, John Guttag, and Emre Kıcıman. 2025. Walk the talk? measuring the faithfulness of large language model explanations. In *International Conference on Learning Representations*.

Tsvetomila Mihaylova and André F. T. Martins. 2019. Scheduled sampling for transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 351–356, Florence, Italy. Association for Computational Linguistics.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning

chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and 1 others. 2025. Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications*, 16(1):9799.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. 2024. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 530–546.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*.

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*.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9935–9960.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, and 1 others. 2025b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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

A Implementation Details

A.1 Datasets

MMLU (Hendrycks et al., 2021a). MMLU is a multiple-choice benchmark designed to evaluate multitask language understanding across academic domains. We use the test set of the benchmark, which contains 14,042 four-choice questions spanning 57 subjects, including STEM, social sciences, and professional fields. Many questions are answerable via knowledge retrieval alone, providing a baseline where CoT may be unnecessary for capable models.

GPQA (Rein et al., 2023). GPQA is a graduate-level multiple-choice Q&A benchmark designed to test scientific reasoning. The dataset contains 448 questions validated by domain experts in biology, physics, and chemistry. Because GPQA questions are intentionally difficult even for highly skilled people, the dataset provides a setting for examining whether reasoning traces materially affect answer distributions on specialized scientific problems.

MATH (Hendrycks et al., 2021b). MATH is an open-ended mathematical reasoning benchmark. We use the test set of this benchmark, consisting of 5,000 competition-level mathematics problems. Each problem includes a step-by-step solution and a final answer, and problems are categorized into seven mathematical subjects and five difficulty levels. Unlike MMLU and GPQA, MATH does not restrict answers to a fixed set of multiple-choice options. CoT is expected to be most beneficial here, as solutions involve chained logical deductions.

Selection logic. These datasets represent a regime where even CoT may be insufficient without specialized knowledge, allowing us to test the upper bound of reasoning augmentation. Also, this selection creates a difficulty gradient (undergraduate $\rightarrow$ competition $\rightarrow$ graduate-expert) and covers qualitatively different reasoning types (knowledge recall $\rightarrow$ formal deduction $\rightarrow$ expert inference).

A.2 Trace Generation

We generate reasoning traces and baseline answers for every model–dataset pair. All runs use the same instruction-tuned LLMs (Qwen 2.5-Instruct and Gemma 3-IT families). Traces are generated once per question and reused across all downstream perturbation and probability-extraction pipelines.

Chain-of-thought traces. For the *with-trace* condition, the model first receives the question together with an instruction to produce explicit step-by-step reasoning. The model is not asked to commit to a final answer in this turn. The resulting text is stored as the reasoning trace S. Next, the question and S are presented again, and the model is instructed to answer using only the final response format required by the task (a single letter for multiple-choice items, or a concise answer string for open-ended items).

Reasoning traces are split with a set of hard rules: (i) "Step N" headers, (ii) numbered list items, (iii) "Statement/Option N" blocks, (iv) paragraph breaks; otherwise the trace is a single step.

Direct (no-trace) baseline. For the *without-trace* condition, the model receives the same question but is instructed not to explain or reason explicitly. It must produce only the required answer format in a single turn. This yields the question-only answer distribution used in the necessity comparisons.

Other details. We use `temperature=0`. No explicit stop sequences are used and generation ends at `max_tokens`. API configurations are as follows.

- MMLU/GPQA: `max_tokens = 1`, `logprobs = True`, `top_logprobs = 5`.
- MATH: `max_tokens = 100`, `logprobs = True`, `top_logprobs` as above.
- Judge calls: `max_tokens= 5` , no logprobs.

Compute. All experiments were conducted on 4$\times$ NVIDIA A100 80GB GPUs over approximately 2,000 A100 GPU-hours of inference compute. The Qwen 2.5 family (640h), Qwen 3 ablation (640h), and Gemma 3 family (470h) account for the majority of the budget. The remaining compute (250h) covers perturbation sensitivity analyses (varying mask ratios in Appendix D.1) and the completeness/ redundancy experiments (Appendix E). Tensor-parallel size was set to 1 for models $\leq$8B, 2 for 14B—27B, and 4 for 32B.

A.3 Measuring Robustness

Robustness is operationalized by applying a family of trace interventions $\mathcal{I}_\lambda$ and comparing the answer token probability distribution before and after perturbation. Our primary intervention is random token masking at rate λ.

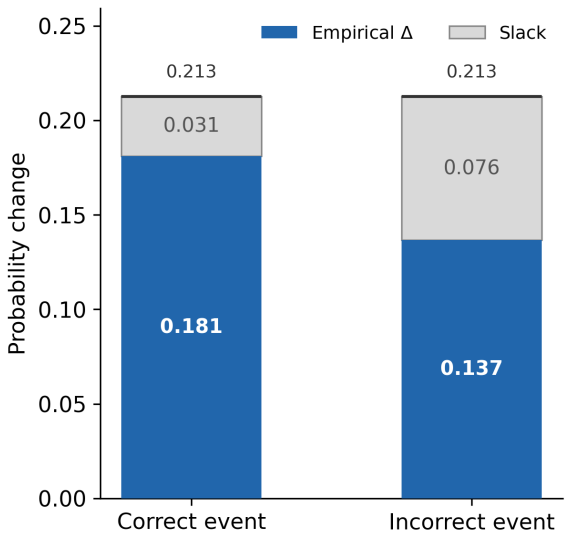

Figure 2: **JS bound slack**. Median empirical Δ and slack for correct vs. incorrect events.

Random token masking. Given a trace S, we construct a perturbed trace S' by independently selecting a fraction λ of trace tokens and replacing each selected token with a dedicated mask symbol (`<redacted_UNK>`). The main analyses fix $\lambda = 10\%$ while we extend the analyses by using $\lambda \in \{5\%, 20\%, 30\%\}$. Masking is applied to the trace tokens only (not to the question), and the perturbed trace is fed forward under the same answer-elicitation process as in trace generation.

Structural trace edits. To study completeness and redundancy in Appendix E, we form three trace variants from the original chain-of-thought output: **CoT**, the unedited reasoning trace; **Dup**, one randomly chosen reasoning step is duplicated in place (redundancy intervention); **SM**, one randomly chosen reasoning step is fully removed.

Each variant is then processed with the same λ-rate token masking, yielding paired distributions $P_m(A \mid X, S)$ and $P_m(A \mid X, S')$ required for robustness computation.

Evaluation target. Robustness is summarized by $-1 \times \mathrm{JS}\big(P_m(A \mid X, S),\ P_m(A \mid X, S')\big)$ computed on the answer distributions induced by the perturbed trace.

Robustness proxy. In Proposition 1, we show that the JS penalty functions as a theoretical upper bound of reasoning instability (i.e., the reverse of robustness). We use $\mathrm{JS}(P^S, P^{S'})$ rather than a single event measure $\Delta(E)$ because it is a distribution-level quantity. That is, one symmetric penalty upper-bounds instability (i.e., lower-bounds robustness) for *every* answer event, whereas $\Delta(E)$ is event-specific, not comparable across items without choosing E, and can miss shifts among distractors even when $\Delta(E^\star)$ is small. We justify this design choice by showing empirically that (i) the inequality holds for the vast majority of cases and that (ii) when measured for correct answers, the inequality becomes tighter.

In our experiment using the entire sample of multiple-choice questions (i.e., MMLU and GPQA), 97.5% satisfy $\Delta \leq \sqrt{2\,\mathrm{JS}}$ (tolerance 10^{-5}). The 2.5% violation is mostly numerical edge cases when $\mathrm{JS} \approx 0$).

We then split the sample into two subsamples: correct events and incorrect events. Figure 2 shows the results. For the correct event, median $\Delta = 0.181$ and slack $= 0.031$; for the incorrect event ($k^\dagger$), median $\Delta = 0.137$ and slack $= 0.076$. The bound is empirically tighter for correct events and the difference is statistically significant at the 1% level. Note that Proposition 1 implies that the upper bound of 0.213 applies universally to all events including both correct and incorrect events.

A.4 Answer Probability Extraction

All empirical quantities are computed from model log-probabilities. Inference requests return the top-200 log-probabilities at each position. Remaining vocabulary mass is assigned uniformly to the complement so that each local distribution sums to one.

Multiple-choice tasks. For multiple choice questions the answer space is the four choice symbols, $\mathcal{A} = \{A, B, C, D\}$. We read the model's next-token distribution at the answer position, extract the probability mass assigned to each choice symbol, and renormalize over $\mathcal{A}$. This yields the direct-condition distribution $P_m(Y \mid X)$ and the trace-conditioned distributions $P_m(Y \mid X, S)$ and $P_m(Y \mid X, S')$, where S' denotes a perturbed reasoning trace. correctness-directed necessity is

$$\mathrm{KL}(\delta_y \| P_m(\cdot \mid X)) - \mathrm{KL}(\delta_y \| P_m(\cdot \mid X, S)).$$

For robustness, we compute the divergence between the choice distributions induced by the original and perturbed traces.

Open-ended tasks. For open-ended questions, answers often consist of multiple tokens, such as mathematical expressions. We therefore estimate answer distributions by teacher forcing (Bengio

et al., 2015; Mihaylova and Martins, 2019). Given a ground-truth answer tokenized as $(t_1, \ldots, t_n)$ using the model's native tokenizer, we condition on the correct prefix $(t_1, \ldots, t_{i-1})$ and extract the model's next-token distribution at each answer position i. In practice, we use the top-200 log probabilities returned by the model and assign any residual probability mass uniformly across the remaining vocabulary. We then compute the token-level score

$$-\log P(t_i \mid X, t_1, \ldots, t_{i-1}),$$

and average it over all answer tokens to obtain a single question-level score. This is equivalent to

$$\mathrm{KL}(\delta_y \| P) = \frac{1}{n} \sum_{i=1}^{n} -\log P(t_i \mid X, t_1, \ldots, t_{i-1}).$$

We apply the same procedure to the direct condition, the original reasoning-trace condition, and the perturbed reasoning-trace condition.

B Statistical Estimations

B.1 Distributions

Figure 3 plots min–max normalized question-level necessity and robustness from Qwen 2.5 (left) and Gemma 3 (right). We summarize the joint distribution of reasoning quality at the question $\times$ model-scale level on GPQA and MMLU. We min–max normalize both metrics within each (dataset, family, question) over the available scales, so both axes lie in $[0, 1]$. For readability, we truncate the measures at the 1% and 99% percentiles within each benchmark–family before normalization. The two measures have a negative Spearman correlation (-0.35 for Qwen 2.5 and -0.23 for Gemma 3), but the correlations are not statistically significant at the 10% level.

In addition, Table 1 divides the questions based on their median values. All four regimes are evenly populated. High-necessity/high-robustness cases correspond to traces that both change the answer distribution and remain stable under perturbation, whereas high-necessity/low-robustness cases correspond to traces that matter but are fragile. Low-necessity/high-robustness cases capture traces that are stable but largely unnecessary, while low-necessity/low-robustness cases capture traces that are neither consequential nor stable. These patterns suggest that necessity and robustness are distinct dimensions that separate qualitatively different reasoning regimes.

Table 1: Four regimes of reasoning quality. Each question-level metric is assigned to one of the quadrants based on their median values.

Regime	Share (%)
High necessity / High robustness	20.3
High necessity / Low robustness	27.2
Low necessity / High robustness	31.5
Low necessity / Low robustness	21.0

Table 2: Pooled quadratic estimation results

Model	Outcome	$\hat{\beta}_2$	p-value
Qwen 2.5	Necessity	-0.077	< 0.001
Qwen 2.5	Robustness	0.001	< 0.001
Gemma 3	Necessity	-0.116	0.063
Gemma 3	Robustness	0.002	0.003

B.2 Mixed-Effects Regression Specification

For model m in family f, let $s_m = \log(\mathrm{size}_m)$ and let $y_{f,d,j,m}$ denote a subtask-level summary for dataset d, subtask j, and outcome y. Subtask is a granular task defined within each dataset (e.g., chemistry in MMLU).

Outcomes. At the item level we use

$$\widehat{\mathrm{Necessity}}_i = \mathrm{KL}(\delta_{y_i} \| P_m^0) - \mathrm{KL}(\delta_{y_i} \| P_m^S),$$
$$\widehat{\mathrm{Robustness}}_i = -\mathrm{JS}(P_m^S, P_m^{S'}),$$

We aggregate the question level metrics at the subtask-model level when there are at least five valid observations.

Mixed effects regression model. Pooled inference within each model family fits

$$y_{f,d,j,m} = \beta_0 + \beta_1 s_m + \beta_2 s_m^2 + u_j + v_d + \varepsilon_{f,d,j,m},$$

where $u_j \sim \mathcal{N}(0, \sigma_u^2)$ is a random intercept for subtask j and $v_d \sim \mathcal{N}(0, \sigma_d^2)$ is a crossed random intercept for dataset d. Parameters are estimated by restricted maximum likelihood (REML) with an L-BFGS optimizer. If the crossed specification fails to converge, we fit random intercepts for subtask only. Inference on β_2 uses the mixed-model covariance with normal approximations. Models are fit separately for each family (Qwen 2.5, Gemma 3) and each outcome.

We test whether β_2 is negative (positive) for necessity (robustness).

Lind–Mehlum endpoint-slope test. Let $\widehat{m}(s) = \hat{\beta}_1 + 2\hat{\beta}_2 s$ denote the derivative of the fitted quadratic mean. At endpoints $s_{\min}$ and $s_{\max}$ we

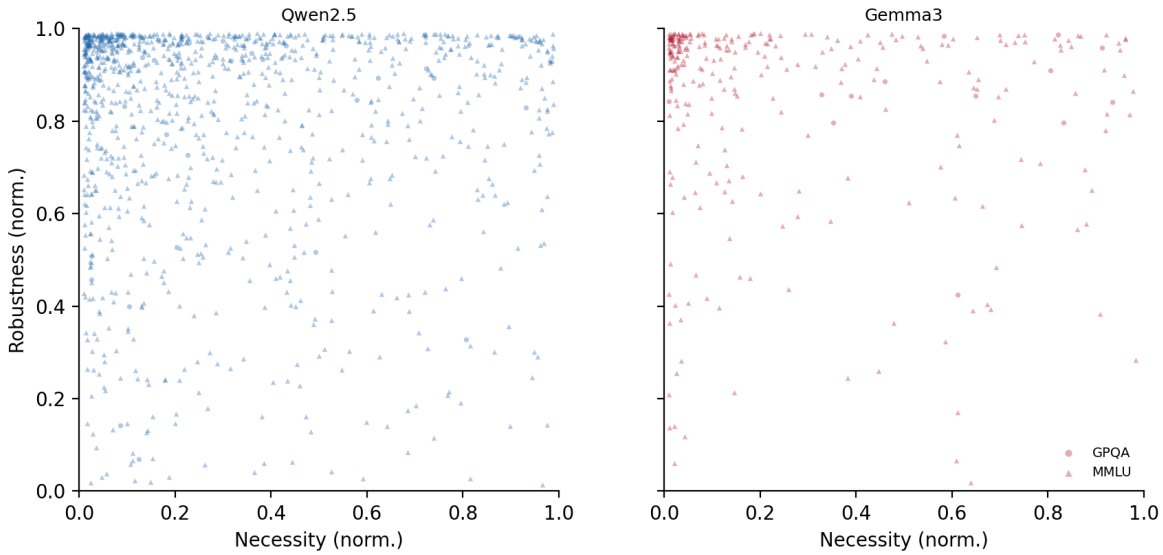

Figure 3: **Distributions**. This figure plots min-max normalized question-level necessity and robustness from Qwen2.5 (left) and Gemma3 (right).

Table 3: Quadratic results by model and dataset

		Necessity		Robustness	
Model	Dataset	$\hat{\beta}_2$	p	$\hat{\beta}_2$	p
Qwen 2.5	GPQA	−0.09	0.05	0.18	0.15
Qwen 2.5	MMLU	−0.06	0.00	0.11	0.00
Qwen 2.5	MATH	−0.21	0.00	0.07	0.24
Gemma 3	GPQA	−0.21	0.12	0.12	0.04
Gemma 3	MMLU	−0.14	0.02	0.14	0.07
Gemma 3	MATH	−0.07	0.06	0.93	0.01

evaluate $\widehat{m}(s_{\min})$ and $\widehat{m}(s_{\max})$ with standard errors from $(\hat{\beta}_1, \hat{\beta}_2)$. For an inverse-U, theory requires a positive slope at the smallest scale and a negative slope at the largest, tested with one-sided t-statistics at each endpoint. For a U-shape the inequalities are reversed.

Estimation results. Table 2 shows the mixed effect regression results. For Qwen 2.5, $\hat{\beta}_2 = -0.077$ ($p < 0.001$) for Necessity and $\hat{\beta}_2 = 0.001$ ($p < 0.001$) for Robustness, with Lind–Mehlum rejection ($p < 0.001$). For Gemma 3, Robustness shows $\hat{\beta}_2 = 0.002$ ($p = 0.003$) with an interior minimum, Necessity has $\hat{\beta}_2 = -0.116$ ($p = 0.063$) with Lind–Mehlum rejection ($p = 0.02$).

Table 3 estimates the mixed effect regressions at the model-dataset level. The results suggest that necessity (robustness) has an inverse-U-shaped (U-shaped) relationship with model scale even at the more granular dataset level.

B.3 Bootstrap Procedure

We perform a subtask bootstrap with 500 replicates.

Resampling unit. Within a family and outcome, let $\mathcal{J}$ denote subtasks. Replicate b draws $|\mathcal{J}|$ subtasks with replacement, concatenates all (subtask, model) rows for the drawn subtasks, and forms a pseudo–scale curve at each s_m by:

1. averaging y within each drawn subtask at s_m;
2. averaging subtask means with equal weight across subtasks present at s_m.

Bootstrap statistic. On each bootstrap curve we fit OLS $y = \alpha_0 + \alpha_1 s + \alpha_2 s^2$. The replicate counts as an interior peak if $\hat{\alpha}_2$ has the predicted sign and the turning point lies in $(s_{\min}, s_{\max})$. We report the fraction of successful replicates with an interior peak among those with at least three distinct scale points. For Qwen 2.5, this fraction is 99.8% (Necessity) and 100% (Robustness); for Gemma 3, 93.6% and 98.4%, respectively.

C Validating Theorem 1's Assumptions

Theorem 1 formalizes one sufficient set of conditions under which necessity and robustness can be non-monotonic in model capability.

C.1 Operationalizing ℓ, σ, r, and u

All expectations below are sample means over items in a benchmark and pooled curves average benchmark-level estimates unless noted otherwise.

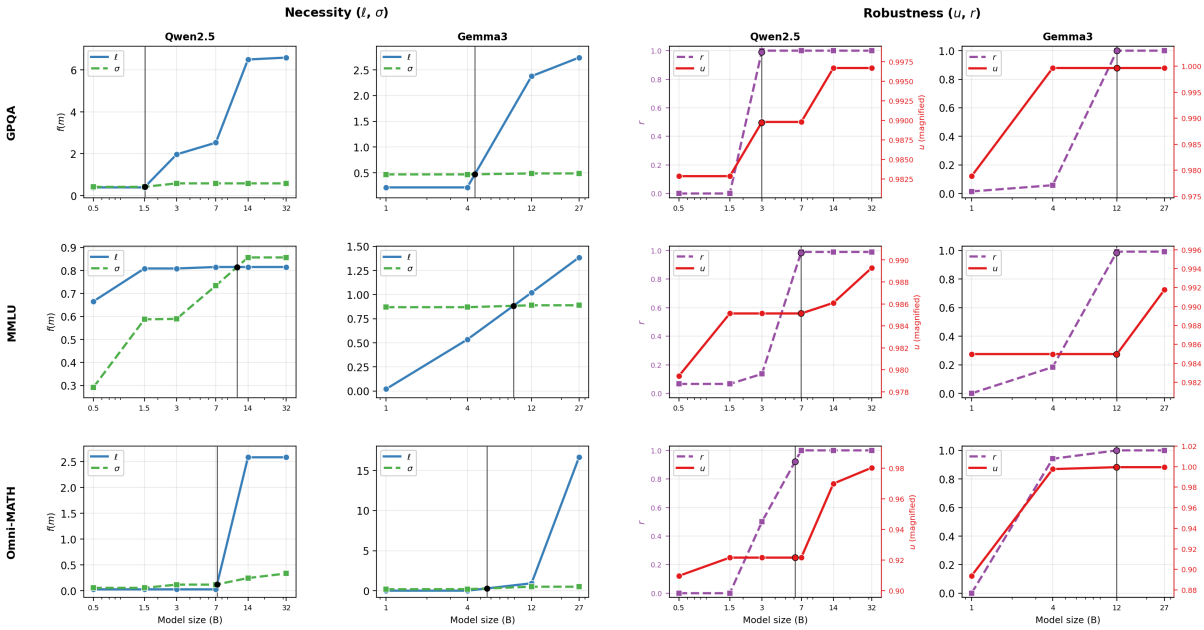

Figure 4: **Single-crossing.** Each row is a benchmark (GPQA, MMLU, MATH); columns are Qwen 2.5 then Gemma 3. Left panels show $\ell(m)$ (solid) and $\sigma(m)$ (dashed). Right panels show $u(m)$ (solid) and $r(m)$ (dashed). ℓ is mean $\mathrm{KL}(P^{\mathrm{CoT}} \| P^{\mathrm{direct}})$; σ is mean $P^{\mathrm{direct}}(y^\star)$; r is $\mathrm{Acc}(\mathrm{CoT}) - \mathrm{Acc}(\mathrm{direct})$; and $u = 1 - \mathbb{E}[\mathrm{JS}]/\ln 2$ under 10% token CoT masking.

Trace-usage ability $\ell(m)$. Theoretical $\ell(m)$ is

$$\ell(m) \;=\; \mathbb{E}_{X,y,S}\big[\mathrm{KL}(P_m^S | P_m^0)\big].$$

A larger $\ell(m)$ means the trace shifts the answer distribution further from the question-only distribution. On multiple-choice tasks we estimate P_m^0 and P_m^S by normalizing reported probability mass on $\{A, B, C, D\}$ at the answer token, then computing $\mathrm{KL}(P_m^S | P_m^0)$ over the four choices. On open-ended items we use the teacher-forced divergence between chain-of-thought and direct conditions, $\mathrm{KL}(P_m^{\mathrm{CoT}} \| P_m^{\mathrm{direct}})$, as the open-ended analogue of the same contrast.

Question self-sufficiency $\sigma(m)$. Self-sufficiency is the expected question-only probability of the correct answer,

$$\sigma(m) \;=\; \mathbb{E}_{X,y}\big[P_m^0(y)\big].$$

For multiple-choice items we take $P_m^0(y)$ from the direct-condition distribution. For open-ended items we use $\exp\big(-\mathrm{KL}(\delta_y \| P_m^0)\big)$, i.e. the teacher-forced probability of the ground-truth answer under the direct question, averaged over items.

Reliance on explicit reasoning $r(m)$. Reliance is the incremental accuracy attributable to the trace, where accuracy is the hard-match rate between the emitted answer and the ground truth. This summarizes how much final performance depends on the trace-based process compared to the question alone.

Recoverability $u(m)$. Recoverability measures how little the answer distribution moves when the trace is perturbed. Let S' be the 10% token-masked trace and $\mathrm{JS}_m = \mathrm{JS}(P_m^S, P_m^{S'})$. Our primary distributional index is

$$u_{\mathrm{JS}}(m) = 1 - \frac{\mathbb{E}[\mathrm{JS}_m]}{\ln 2},$$

so $u_{\mathrm{JS}} = 1$ indicates no shift and smaller values indicate lower robustness.

C.2 Empirical Curves

We plot ℓ, σ, r, and u with respect to model size m (log-scaled).

Figure 4 illustrates the results. Consistent with the assumptions in Theorem 1, ℓ, σ, r, and u weakly increases with model scale m. Also, ℓ and σ, and r and u single-cross with each other, respectively.

D Additional Results

D.1 Different Masking Ratios

Our baseline analysis masks 10% of the tokens in the reasoning trace. Using Qwen 2.5 on the

Table 4: **Necessity and robustness**. Each cell is the average of per-question **min–max normalized scores**, with 99% CI in brackets.

		GPQA		MMLU		MATH	
Family	Scale	Necessity↑	Robustness↑	Necessity↑	Robustness↑	Necessity↑	Robustness↑
Qwen 2.5	0.5B	0.28 [0.02, 0.55]	0.91 [0.88, 0.94]	0.11 [0.09, 0.12]	0.89 [0.88, 0.89]	0.50 [0.35, 0.66]	0.70 [0.60, 0.80]
	1.5B	0.28 [0.08, 0.48]	0.55 [0.46, 0.64]	0.34 [0.32, 0.36]	0.44 [0.43, 0.46]	0.49 [0.39, 0.59]	0.61 [0.49, 0.72]
	3B	0.46 [0.24, 0.69]	0.68 [0.60, 0.77]	0.69 [0.67, 0.71]	0.69 [0.67, 0.71]	**0.76 [0.69, 0.84]**	0.64 [0.55, 0.73]
	7B	**0.49 [0.26, 0.72]**	0.76 [0.68, 0.84]	**0.73 [0.71, 0.74]**	0.84 [0.82, 0.85]	0.71 [0.64, 0.78]	0.42 [0.32, 0.51]
	14B	0.42 [0.20, 0.64]	0.89 [0.84, 0.95]	0.57 [0.55, 0.59]	0.92 [0.91, 0.93]	0.65 [0.58, 0.73]	0.77 [0.69, 0.86]
	32B	0.17 [0.03, 0.30]	**0.99 [0.98, 1.00]**	0.46 [0.44, 0.48]	**0.99 [0.99, 1.00]**	0.36 [0.28, 0.44]	**0.98 [0.95, 1.00]**
Gemma 3	1B	0.25 [-0.26, 0.77]	0.87 [0.82, 0.92]	0.26 [0.23, 0.29]	0.82 [0.80, 0.83]	0.49 [0.46, 0.53]	**0.86 [0.84, 0.88]**
	4B	**0.97 [0.88, 1.05]**	0.61 [0.50, 0.72]	**0.79 [0.75, 0.82]**	0.48 [0.46, 0.51]	0.71 [0.69, 0.74]	0.27 [0.24, 0.30]
	12B	0.26 [-0.22, 0.75]	0.34 [0.23, 0.45]	0.44 [0.39, 0.48]	0.48 [0.46, 0.50]	**0.76 [0.74, 0.77]**	0.64 [0.60, 0.67]
	27B	0.24 [-0.27, 0.74]	**0.99 [0.98, 1.00]**	0.17 [0.14, 0.20]	**1.00 [1.00, 1.00]**	0.30 [0.27, 0.32]	0.82 [0.80, 0.84]

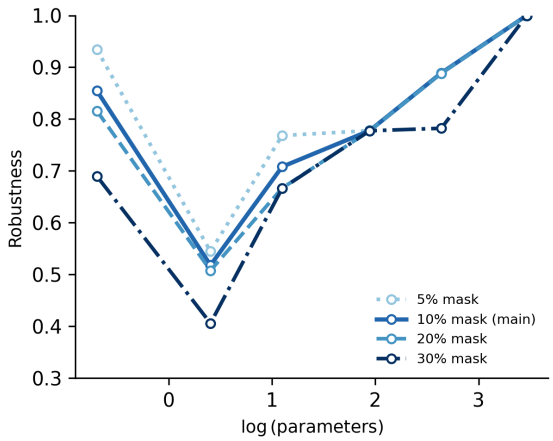

Figure 5: **Robustness measured changing the masking ratio**. We use Qwen 2.5 on the GPQA benchmark.

GPQA benchmark, we repeat the analysis using masking ratios of 5%, 10%, 20%, and 30%. In Figure 5, across all masking ratios, robustness is lowest for intermediate-scale models and higher for both smaller and larger models. The magnitude of the JS penalty increases mechanically as the perturbation becomes stronger, but the relative ordering across model sizes is largely preserved.

D.2 Subtask-Level Data

In addition to the benchmark-model level illustrations in Figure 1, we plot subtask-level necessity and robustness in Figure 6. We find evidence consistent with necessity having a inverse-u shaped relationship with model scale and robustness having a u shaped relationship with model scale. The underlying data of this figure is in Table 4.

D.3 Conditioning on Positive Necessity

In addition to the main experiment, we condition on necessity being positive and examine whether necessity and robustness still maintain their non-monotonic relationship with model scale. Intuitively, we restrict the sample to where reasoning traces help the model to improve directional accuracy. The results are displayed in Figure 7. The non-monotonic relationships remain robust.

D.4 Using a Reasoning-Tuned Model

Qwen 2.5 and Gemma 3 provide different scales within a single family (six and four sizes, respectively), spanning roughly two orders of magnitude in parameter count. This provides us with an ideal setting to empirically test the inverse-U and U-shape predictions of Theorem 1. In addition because the two modes share the same weights, the within-model paired comparison is free from confounds introduced by comparing different checkpoints (e.g., R1 vs. its base, or Instruct vs. Thinking variants released as separate weights).

We use instruction-tuned but not reasoning-specialized models in the main experiment for two reasons.

First, our framework requires that the question-only distribution $P_m(A|X)$ be well-defined. Necessity computation is meaningful only when the model produces an interpretable answer distribution under the no-trace condition. Reasoning-specialized models are post-trained to always emit explicit reasoning, and even when forced to answer directly their next-token distribution tends to concentrate on conversational filler (e.g., "Okay", "Alright") rather than the answer token space. The induced $P_m(A|X)$ is thus not likely a model belief over answers, making necessity uninterpretable.

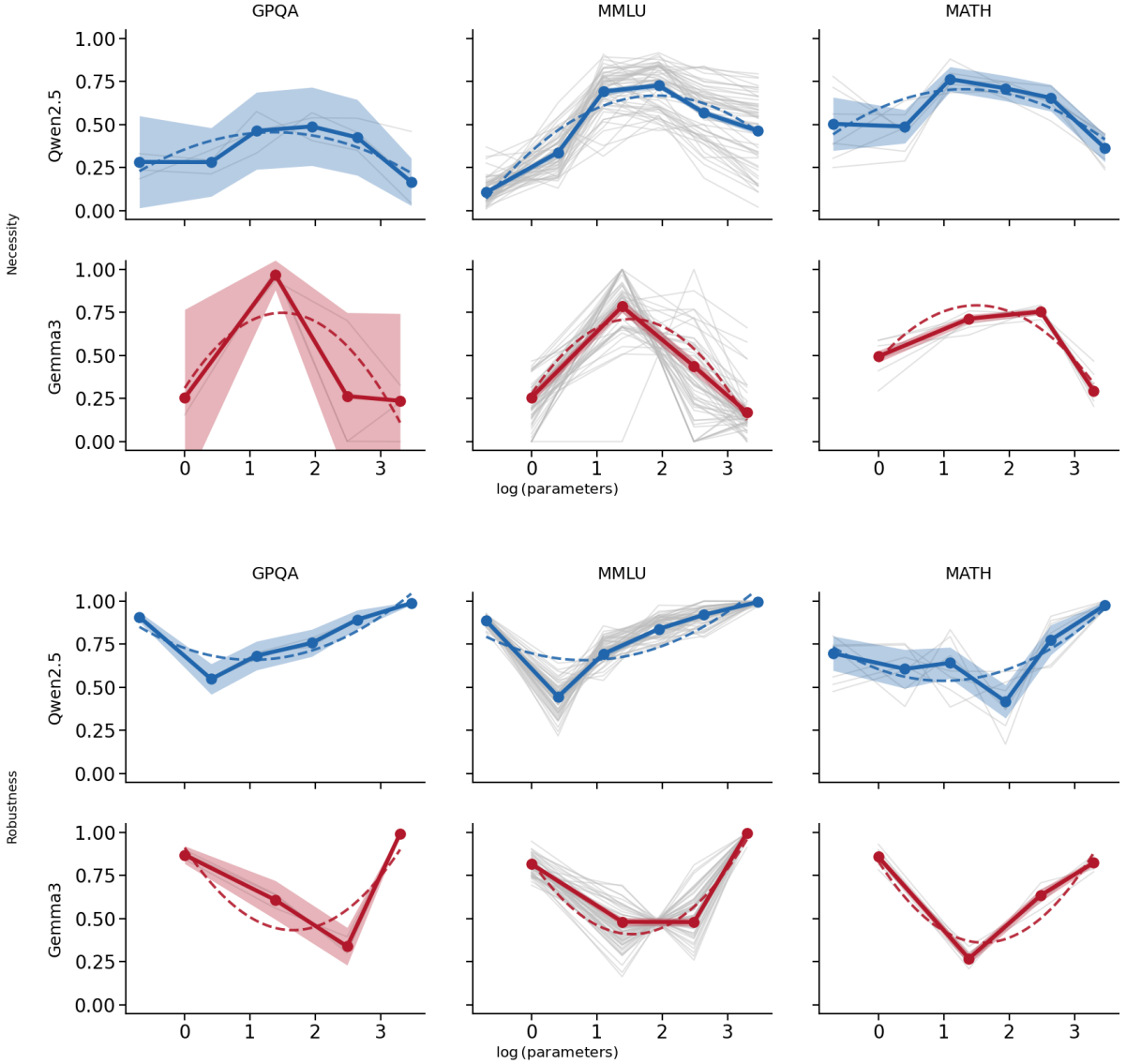

Figure 6: **Necessity and robustness at the subtask level**. Per-question min–max normalized necessity (robustness) vs. log model size. Shaded bands are 99% confidence intervals. Grey lines denote subtask-level measures.

Second, current reasoning-specialized models are increasingly released as mode-separated weights (e.g., DeepSeek V3/R1) without a counterpart, thereby making within-model comparison challenging.

Nevertheless, to test robustness of our results, we repeat the experiment on Qwen 3, one of the few model families that support both thinking and non-thinking operation under identical parameters. In a non-thinking mode, the model emits the answer directly without a thinking trace, so the no-trace distribution concentrates on the answer tokens and remains well-defined.

The results are reported in Figure 8. We find that non-monotonic relationships between model scale and reasoning quality persist when explicit reasoning is activated.

E Linguistic Attributes

E.1 Completeness Experiments

The completeness experiment evaluates whether reasoning traces become less robust when an inferential step is removed. We operationalize incompleteness by constructing a step-removed version of each reasoning trace and then recomputing JS penalty under the same masking procedure used in the main robustness analysis. Let S denote the original reasoning trace and let S^{-} denote the trace

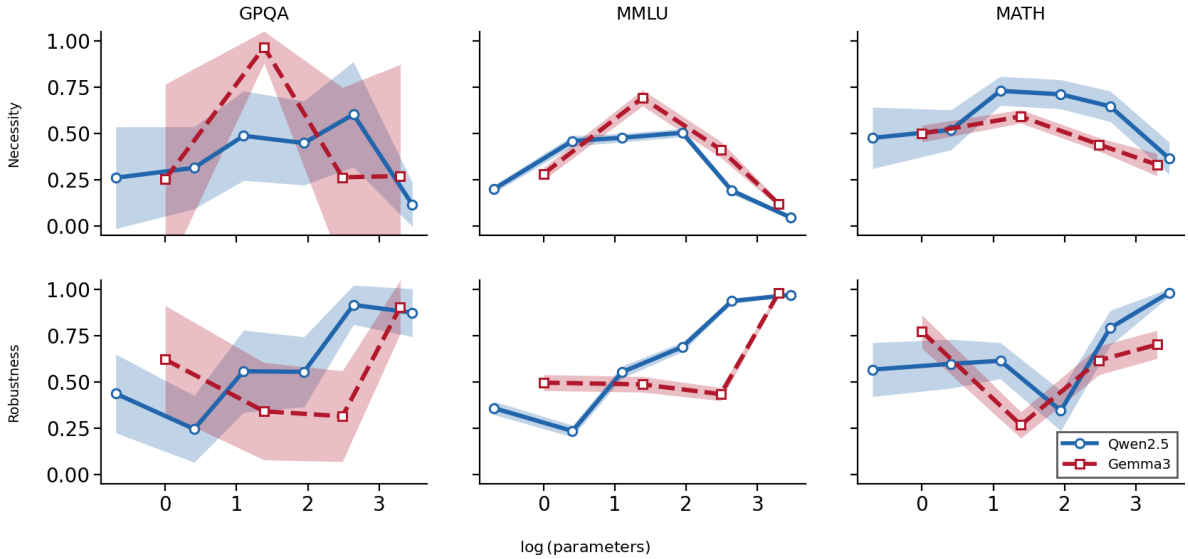

Figure 7: **Necessity and robustness aggregated at the model-benchmark level when necessity is positive**. Per-question min–max normalized necessity (top) and robustness (bottom) vs. log model size on shape-selected questions from GPQA, MMLU, and MATH. Samples are restricted to when necessity values are positive. Shaded bands are 99% confidence intervals across questions.

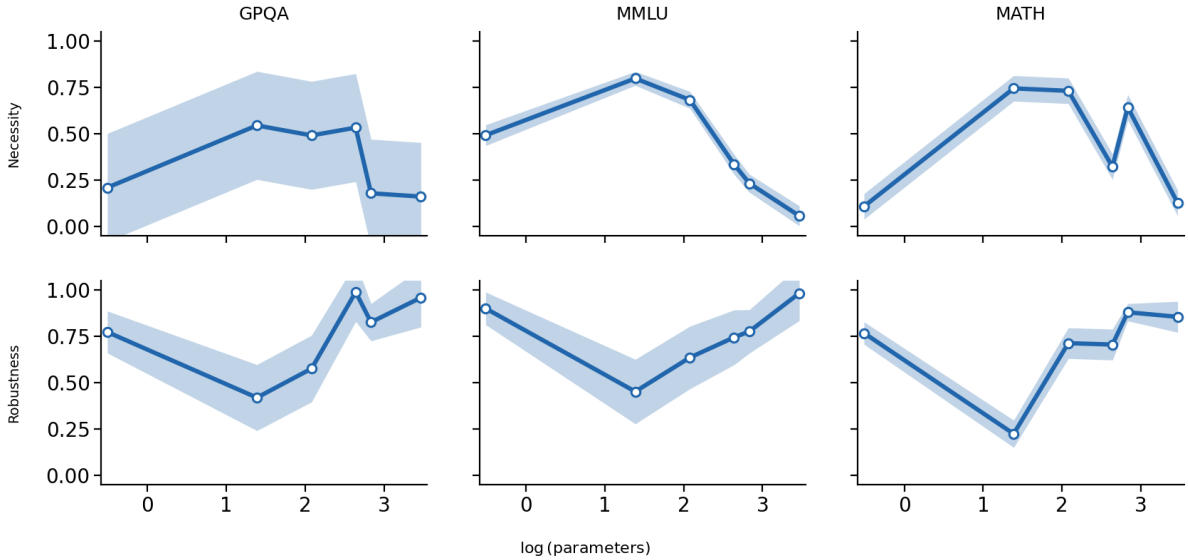

Figure 8: **Qwen 3**. We use Qwen 3 instead of Qwen 2.5. Per-question min–max normalized necessity (top) and robustness (bottom) vs. log model size on shape-selected questions from GPQA, MMLU, and MATH. Samples are restricted to when necessity values are positive. Shaded bands are 99% confidence intervals across questions.

after one reasoning step has been removed. We compare

$$\mathrm{JS}\Big(P_m(A \mid X, S), P_m(A \mid X, \widetilde{S})\Big)$$

and

$$\mathrm{JS}\Big(P_m(A \mid X, S^-), P_m(A \mid X, \widetilde{S}^-)\Big),$$

where $\widetilde{S}$ and $\widetilde{S}^-$ denote masked versions of the original and step-removed traces, respectively.

The theoretical prediction (see Appendix G.3) is that step removal should reduce robustness when the removed step contains information that the model uses. In that case, the trace becomes less complete, and the model's answer distribution be-

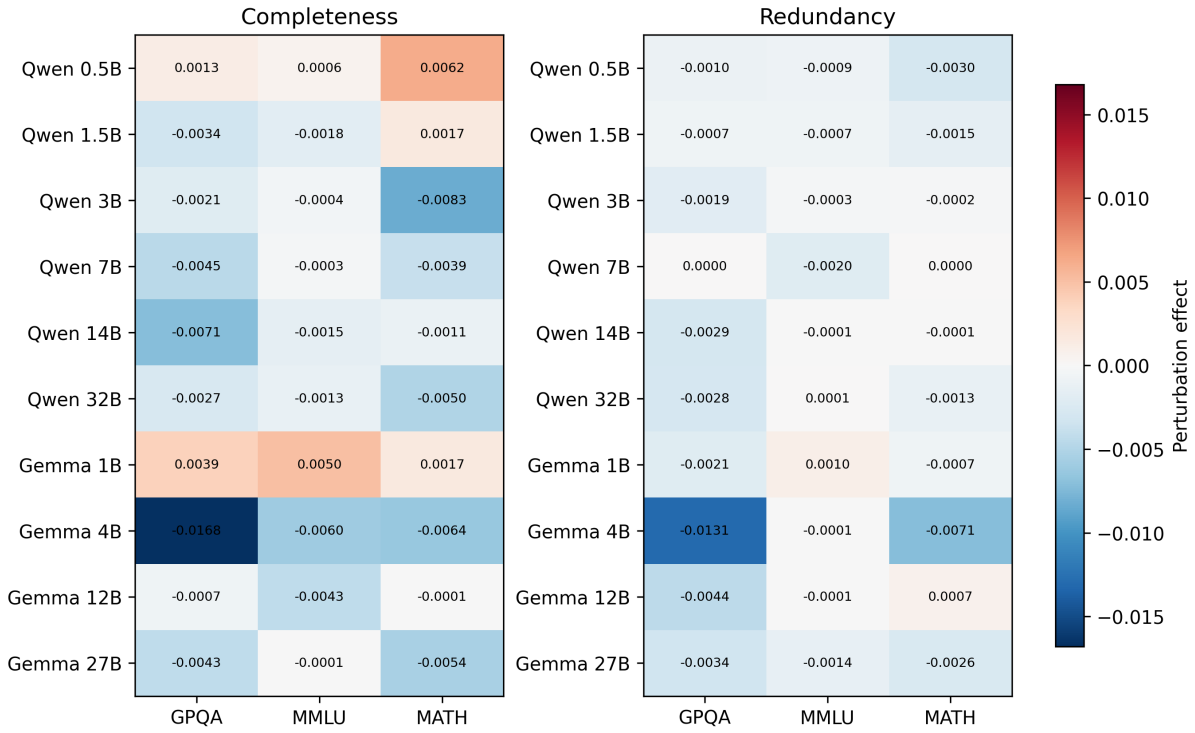

Figure 9: **Completeness and redundancy perturbation effects**. Left panel reports the signed completeness effect $JS(CoT) - JS(SM)$, and right panel reports the signed redundancy effect $JS(CoT) - JS(Dup)$.

comes more sensitive to additional perturbations. Therefore, the empirical implication is

$$\mathrm{JS}(S^{-}) > \mathrm{JS}(S).$$

Figure 9 visualizes the results. For larger models in general, the step-removed trace produces higher JS penalty than the original trace across all masking ratios. This suggests that, for models with sufficient capacity, individual reasoning steps are informative and removing them makes the trace less complete. Thus, step removal increases the model's sensitivity to perturbations. However, the evidence is weaker for smaller models. This is consistent with smaller models not fully exploiting the reasoning trace, so deleting one step does not necessarily create a additional loss of information.

Overall, the experiment supports the interpretation that completeness can be understood as a property of robustness: complete traces are stable under admissible perturbations.

E.2 Redundancy Experiments

We evaluate whether duplicating reasoning steps reduces robustness. We construct a duplicated-step trace S^{+} by repeating one reasoning step in the original chain-of-thought trace. We then compare the JS of the duplicated trace against that of the original trace under the same token-masking.

The theoretical prediction is that if each duplicated step creates an additional location at which perturbations can affect the induced answer distribution, then the accumulated upper bound of JS penalty may increase with the number of redundant steps (see Appendix G.4). Under this channel,

$$\mathrm{JS}(S^{+}) > \mathrm{JS}(S).$$

The empirical results in Figure 9 generally support this prediction. Duplicated traces generally have higher JS penalty than the original traces. In several cases, however, duplication slightly improves robustness. An interpretation is that duplicated reasoning repeats an informative trace and makes the relevant inference easier for the model to recover after local token masking.

F Prompts and Templates

All experiments use temperature 0 unless noted otherwise. Placeholders are filled at runtime: `{subject}`, `{question}`, `{choice_A}`–`{choice_D}` (MMLU and GPQA), or `{problem}` (MATH).

F.1 Trace-Generation Prompts

MMLU/GPQA — Chain-of-Thought Two-turn dialogue: (i) trace generation, then (ii) answer extraction (Section F.2).

MMLU/GPQA CoT trace

The following is a multiple choice question about {subject}.
Question: {question} A. {choice_A} B. {choice_B} C. {choice_C} D. {choice_D}
Think step by step, analyzing the logic behind each option. Write the reasoning process as Step 1: ~~, Step 2: ~~ … Step n: ~~
RULES:

- You MAY reason about which options are more or less plausible, and why.
- You MAY state, in theory, what the correct concept or principle is.
- **STOP** before explicitly selecting or labeling any option as the final answer.
- Do NOT write phrases like "therefore the answer is", "the correct option is", "this matches option X", or any equivalent conclusion that names a letter as the answer.
- End your response as if handing off your reasoning to someone else to make the final selection.

MATH — Chain-of-Thought. Math problem solving with trace.

Math CoT trace

Solve the following math problem step by step.
Problem: {problem}
Think step by step, analyzing the problem carefully.
RULES:

- You MAY reason about the problem and show your work.
- **STOP** before stating your final answer.
- Do NOT write phrases like "therefore the answer is" or any equivalent conclusion.
- End your response as if handing off your reasoning to someone else to state the final answer.

[Same as above, plus:]
Write the reasoning process as Step 1: ~~, Step 2: ~~ … Step n: ~~

MATH — Direct. No trace; single-turn answer only.

Math Direct trace (no reasoning)

Solve the following math problem. Do not explain or show any work. Answer with ONLY the final answer (a number, expression, or short mathematical notation). Nothing else.
Problem: {problem}
Answer:

F.2 Answer-Extraction Prompts

MMLU/GPQA. After CoT trace, we request a single-letter answer and read log-probabilities over {A,B,C,D} from the first generated token.

MMLU/GPQA answer extraction (Direct)

Now answer with only a single letter (A, B, C, or D). The answer is

MATH. Open-ended answers; we extract a short final answer (max 100 tokens) with logprobs, then optionally judge equivalence.

MATH answer extraction (CoT & paraphrase)

Now state ONLY the final answer (a number, expression, or short mathematical notation). Nothing else.
Answer:

Math answer equivalence judge (accuracy metric)

You are a math answer equivalence checker. Determine if the given answer is mathematically equivalent to the correct answer.
Correct answer: {correct_answer} Given answer: {given_answer}
Are these two answers mathematically equivalent? Answer with ONLY "yes" or "no".

G Theoretical Material

G.1 Proof of Proposition 1

Proof of Proposition 1. Fix (X, S) and let $S' \sim \mathcal{I}_\lambda(S)$. Define

$$p := P_m(A \mid X, S),$$
$$q := P_m(A \mid X, S').$$

For any event $E \subseteq \mathcal{A}$, the definition of total variation distance implies

$$|p(E) - q(E)| \leq \mathrm{TV}(p, q).$$

We next relate total variation distance to Jensen–Shannon divergence. Let

$$\bar{p} := \frac{1}{2}(p + q).$$

By Pinsker's inequality,

$$\mathrm{KL}(p\|\bar{p}) \geq 2\mathrm{TV}(p, \bar{p})^2,$$
$$\mathrm{KL}(q\|\bar{p}) \geq 2\mathrm{TV}(q, \bar{p})^2. \tag{3}$$

Since $\bar{p} = (p + q)/2$,

$$\mathrm{TV}(p, \bar{p}) = \mathrm{TV}(q, \bar{p}) = \frac{1}{2}\mathrm{TV}(p, q).$$

Therefore,

$$\begin{aligned}
\mathrm{JS}(p, q) &= \frac{1}{2}\mathrm{KL}(p\|\bar{p}) + \frac{1}{2}\mathrm{KL}(q\|\bar{p}) \\
&\geq \frac{1}{2} \cdot 2\left(\frac{1}{2}\mathrm{TV}(p, q)\right)^2 + \frac{1}{2} \cdot 2\left(\frac{1}{2}\mathrm{TV}(p, q)\right)^2 \\
&= \frac{1}{2}\mathrm{TV}(p, q)^2. \qquad (4)
\end{aligned}$$

Rearranging gives

$$\mathrm{TV}(p, q) \leq \sqrt{2\mathrm{JS}(p, q)}.$$

Combining the two inequalities yields

$$|p(E) - q(E)| \leq \sqrt{2\mathrm{JS}(p, q)}. \tag{5}$$

Substituting back the model-induced answer distributions gives

$$\left|P_m(E \mid X, S) - P_m(E \mid X, S')\right| \leq \sqrt{2\mathrm{JS}(P_m(A \mid X, S), P_m(A \mid X, S'))}. \tag{6}$$

Taking expectations over $S' \sim \mathcal{I}_\lambda(S)$,

$$\Delta_{m,\lambda}(X, S; E) \leq \mathbb{E}_{S' \sim \mathcal{I}_\lambda(S)} \left[\sqrt{2\mathrm{JS}(P_m(A \mid X, S), P_m(A \mid X, S'))}\right]. \tag{7}$$

For the second inequality, apply Jensen's inequality to the concave function $z \mapsto \sqrt{z}$:

$$\mathbb{E}[\sqrt{Z}] \leq \sqrt{\mathbb{E}[Z]}.$$

Taking

$$Z = 2\mathrm{JS}\big(P_m(A \mid X, S), P_m(A \mid X, S')\big),$$

we obtain

$$\begin{aligned} \mathbb{E}_{S'}[\sqrt{Z}] &\leq \sqrt{2\mathbb{E}_{S' \sim \mathcal{I}_\lambda(S)} \left[\mathrm{JS}(P_m(A \mid X, S), P_m(A \mid X, S'))\right]} \\ &= \sqrt{2\mathcal{P}_{m,\lambda}(X, S)}. \end{aligned} \tag{8}$$

This proves the claim. □

G.2 Proof of Theorem 1

Proof of Theorem 1. Fix (X, S, λ). We first consider necessity. Let

$$P_m^0 = P_m(A \mid X), \qquad P_m^S = P_m(A \mid X, S).$$

Let $T(\cdot \mid X, S)$ denote the answer distribution obtained when the information encoded in the reasoning trace is fully extracted and used. Let $B(\cdot \mid X, S)$ denote the baseline answer distribution that would arise without effective use of the trace. Suppose that the question-only distribution can be written as

$$P_m(A \mid X) = (1 - \sigma(m))B(A \mid X, S) + \sigma(m)T(A \mid X, S), \tag{9}$$

where $\sigma(m) \in [0, 1]$ captures prompt self-sufficiency.

Suppose also that conditioning on the trace moves the model toward the fully trace-informed distribution according to

$$P_m(A \mid X, S) = P_m(A \mid X) + \ell(m)\left[T(A \mid X, S) - P_m(A \mid X)\right], \tag{10}$$

where $\ell(m) \in [0, 1]$ captures trace-usage ability. Subtracting $P_m(A \mid X)$ from both sides gives

$$P_m(A \mid X, S) - P_m(A \mid X) = \ell(m)\left[T(A \mid X, S) - P_m(A \mid X)\right]. \tag{11}$$

Using Equation (9),

$$T(A \mid X, S) - P_m(A \mid X) = (1 - \sigma(m))\left[T(A \mid X, S) - B(A \mid X, S)\right]. \tag{12}$$

Therefore,

$$P_m(A \mid X, S) - P_m(A \mid X) = \ell(m)(1 - \sigma(m))\left[T(A \mid X, S) - B(A \mid X, S)\right]. \tag{13}$$

Equation (13) shows that the trace-induced movement in the answer distribution is governed by

$$h_N(m) := \ell(m)(1 - \sigma(m)). \tag{14}$$

For fixed (X, S), the KL gain

$$\mathcal{N}_m(X, S) := \mathrm{KL}\big(P_m(A) \,\|\, P_m^0\big) - \mathrm{KL}\big(P_m(A) \,\|\, P_m^S\big) \tag{15}$$

is order-preserving in the scalar magnitude of the trace-induced update in Equation (13). That is, there exists a differentiable increasing function G such that

$$\mathcal{N}_m(X, S) = G\big(h_N(m)\big), \qquad G'(z) > 0. \tag{16}$$

Thus, along the fixed update direction induced by the trace, larger movements toward the trace-informed distribution correspond to larger KL necessity gains.

On the interior where $\ell(m) > 0$ and $1 - \sigma(m) > 0$,

$$\frac{d}{dm} \log h_N(m) = \frac{\ell'(m)}{\ell(m)} - \frac{\sigma'(m)}{1 - \sigma(m)}. \tag{17}$$

By the single-crossing condition, there exists a unique $m_N^\star$ such that

$$\frac{\ell'(m_N^\star)}{\ell(m_N^\star)} = \frac{\sigma'(m_N^\star)}{1 - \sigma(m_N^\star)}, \tag{18}$$

with the left-hand side larger before $m_N^\star$ and smaller after $m_N^\star$. Thus,

$$\begin{aligned} h_N'(m) &> 0 \quad \text{for } m < m_N^\star, \\ h_N'(m) &= 0 \quad \text{for } m = m_N^\star, \\ h_N'(m) &< 0 \quad \text{for } m > m_N^\star. \end{aligned} \tag{19}$$

Therefore $h_N(m)$ is single-peaked. Since $\mathcal{N}_m(X, S) = G(h_N(m))$ and G is increasing, the KL-based necessity measure in Equation (15) is also single-peaked in model capability m.

We next consider robustness. Recall that robustness is measured through the negative of the JS instability penalty. Define

$$\mathcal{P}_{m,\lambda}(X, S) := \mathbb{E}_{S' \sim I_\lambda(S)} \left[\mathrm{JS}\left(P_m(A \mid X, S), P_m(A \mid X, S')\right) \right]. \tag{20}$$

By the structure imposed in the theorem, write

$$\mathcal{P}_{m,\lambda}(X, S) = c_\lambda(X, S) r(m)(1 - u(m)), \tag{21}$$

where $c_\lambda(X, S) \geq 0$ does not vary with model capability, $r(m)$ captures reliance on explicit intermediate reasoning, and $u(m)$ captures recoverability under perturbation.

Since $c_\lambda(X, S)$ is fixed with respect to m, the scale pattern of the JS instability penalty is governed by

$$h_I(m) := r(m)(1 - u(m)). \tag{22}$$

On the interior where $r(m) > 0$ and $1 - u(m) > 0$,

$$\frac{d}{dm} \log h_I(m) = \frac{r'(m)}{r(m)} - \frac{u'(m)}{1 - u(m)}. \tag{23}$$

By the single-crossing condition, there exists a unique $m_I^\star$ such that

$$\frac{r'(m_I^\star)}{r(m_I^\star)} = \frac{u'(m_I^\star)}{1 - u(m_I^\star)}, \tag{24}$$

with the left-hand side larger before $m_I^\star$ and smaller after $m_I^\star$. Therefore,

$$\begin{aligned} h_I'(m) &> 0 \quad \text{for } m < m_I^\star, \\ h_I'(m) &= 0 \quad \text{for } m = m_I^\star, \\ h_I'(m) &< 0 \quad \text{for } m > m_I^\star. \end{aligned} \tag{25}$$

Thus $h_I(m)$ is single-peaked, and hence the JS instability penalty $\mathcal{P}_{m,\lambda}(X, S)$ is single-peaked in model capability m.

Since robustness is defined as the negative JS penalty,

$$\mathcal{R}_{m,\lambda}(X, S) := -\mathcal{P}_{m,\lambda}(X, S), \tag{26}$$

robustness is U-shaped in model capability. Combining the two parts proves that necessity is single-peaked, while robustness is U-shaped, under the stated single-crossing condition. □

G.3 Completeness and Perfect Robustness

We formalize completeness as a stability property of the answer distribution under admissible perturbations of the reasoning trace.

Definition 3 (λ-completeness)**.** For a fixed model m, question X, trace S, and perturbation level λ, the trace S is λ-complete if

$$P_m(A \mid X, S') = P_m(A \mid X, S) \qquad \text{for all } S' \sim \mathcal{I}_\lambda(S). \tag{27}$$

Proposition 2 (Completeness implies perfect robustness)**.** *If S is λ-complete for (m, X), then for every event $E \subseteq \mathcal{A}$,*

$$\Delta_{m,\lambda}(X, S; E) = 0 \qquad \textit{and} \qquad \mathcal{P}_{m,\lambda}(X, S) = 0. \tag{28}$$

Proof. Because S is λ-complete,

$$P_m(A \mid X, S') = P_m(A \mid X, S)$$

for every admissible perturbation $S' \sim \mathcal{I}_\lambda(S)$. Hence, for any event $E \subseteq \mathcal{A}$,

$$P_m(E \mid X, S') = P_m(E \mid X, S).$$

Therefore,

$$\big|P_m(E \mid X, S) - P_m(E \mid X, S')\big| = 0$$

for every S', and taking expectations over $S' \sim \mathcal{I}_\lambda(S)$ gives

$$\Delta_{m,\lambda}(X, S; E) = 0. \tag{29}$$

Moreover, the Jensen–Shannon divergence between two identical distributions is zero. Thus,

$$\mathrm{JS}\big(P_m(A \mid X, S), P_m(A \mid X, S')\big) = 0$$

for every admissible perturbation S', and therefore

$$\begin{aligned} \mathcal{P}_{m,\lambda}(X, S) &= \mathbb{E}_{S' \sim \mathcal{I}_\lambda(S)} \left[\mathrm{JS}\big(P_m(A \mid X, S), P_m(A \mid X, S')\big)\right] \\ &= 0. \end{aligned} \tag{30}$$

This is a direct consequence of the definition of λ-completeness. □

G.4 Redundancy and a Worst-Case Bound

Redundancy means that the trace contains repeated or additional reasoning steps that are not necessary for reaching the conclusion. The following result shows that, in the worst case, additional perturbable locations can increase the upper bound on event-level instability (i.e., the lower bound on robustness).

Proposition 3. *Let $S = (S_1, \ldots, S_n)$ be a reasoning trace decomposed into n reasoning components. For each $k = 1, \ldots, n$, let $H_{k,\lambda}$ denote the trace obtained after perturbing the first k components, with $H_{0,\lambda} = S$. Suppose that each component-wise perturbation satisfies*

$$\mathbb{E}\left[\mathrm{TV}(P_m(A \mid X, H_{k-1,\lambda}), P_m(A \mid X, H_{k,\lambda}))\right] \leq \delta(\lambda). \tag{31}$$

Then, for any event $E \subseteq \mathcal{A}$,

$$\mathbb{E}\left[|P_m(E \mid X, S) - P_m(E \mid X, H_{n,\lambda})|\right] \leq n\delta(\lambda). \tag{32}$$

Proof. Let

$$P_k := P_m(A \mid X, H_{k,\lambda}),$$
$$P_0 := P_m(A \mid X, S).$$

For any event $E \subseteq \mathcal{A}$,

$$|P_m(E \mid X, S) - P_m(E \mid X, H_{n,\lambda})| \leq \mathrm{TV}(P_0, P_n).$$

By the triangle inequality for total variation distance,

$$\mathrm{TV}(P_0, P_n) \leq \sum_{k=1}^{n} \mathrm{TV}(P_{k-1}, P_k). \tag{33}$$

Taking expectations gives

$$\mathbb{E}[\mathrm{TV}(P_0, P_n)] \leq \sum_{k=1}^{n} \mathbb{E}[\mathrm{TV}(P_{k-1}, P_k)]. \tag{34}$$

By assumption,

$$\mathbb{E}[\mathrm{TV}(P_{k-1}, P_k)] \leq \delta(\lambda) \qquad \text{for all } k = 1, \ldots, n. \tag{35}$$

Therefore,

$$\mathbb{E}[\mathrm{TV}(P_0, P_n)] \leq n\delta(\lambda).$$

Combining the displays yields

$$\mathbb{E}\left[|P_m(E \mid X, S) - P_m(E \mid X, H_{n,\lambda})|\right] \leq n\delta(\lambda). \tag{36}$$

□

This does not imply that redundancy must always reduce robustness. Rather, it shows that when redundant reasoning introduces additional perturbable locations, the maximum accumulated instability (i.e., lower bound of robustness) can grow with the number of such locations. We also assume that, for simplicity, component-wise TV bound $\delta(\lambda)$ is uniform across k. However, we acknowledge that different reasoning steps may differ in perturbability.