Is That Your Final Answer? Test-Time Scaling Improves Selective Question Answering

William Jurayj **Jeffrey Cheng** **Benjamin Van Durme**
Johns Hopkins University
{wjurayj1,jcheng71,vandurme}@jhu.edu

Abstract

Scaling the test-time compute of large language models has demonstrated impressive performance on reasoning benchmarks. However, existing evaluations of test-time scaling make the strong assumption that a reasoning system should always give an answer to any question provided. This overlooks concerns about whether a model is *confident* in its answer, and whether it is appropriate to always provide a response. To address these concerns, we extract confidence scores during reasoning for thresholding model responses. We find that increasing compute budget at inference time not only helps models answer more questions correctly, but also increases confidence in correct responses. We then extend the current paradigm of *zero-risk* responses during evaluation by considering settings with non-zero levels of response risk, and suggest a recipe for reporting evaluations under these settings.[1]

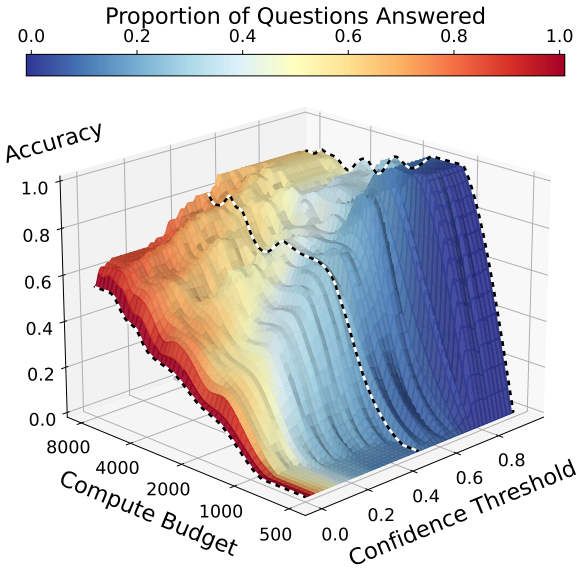


Figure 1: **DeepSeek R1-32B's accuracy is a function of compute budget and confidence threshold.** Increased confidence thresholds generally yield increased accuracy at the cost of response rate, while increased compute budgets sometimes decrease accuracy as they increase response rate. The vertical axis measures the accuracy of answered questions at a compute budget and confidence threshold. Color indicates the proportion of questions that are answered; in redder regions, the model is more likely to answer, whereas in bluer regions the model is less likely to answer.

1 Introduction

Scaling up language model inference-time compute using lengthy chains of thought has delivered impressive results on mathematical reasoning benchmarks that resisted training compute scaling (DeepSeek-AI et al., 2025; Muennighoff et al., 2025). These results, however, are reported in the zero-risk response setting: with no penalties for incorrect answers, the system always guesses even when it is not confident in its answer. In practice, this behavior is not always desirable.

Many question answering settings associate incorrect answers with measurable costs, ranging from low-risk responses found in game shows (Ferrucci et al., 2010) to high-stakes responses that can alter people's lives (Northpointe, 2017). *Selective question answering* addresses these challenges by allowing a model to refrain from answering questions which it might answer incorrectly (Kamath et al., 2020). This requires a selection function, which considers risk tolerance, coverage goals, and candidate answer confidence to decide whether a prediction should be given (Geifman and El-Yaniv, 2017). Knowing when not to answer is a critical quality for systems to collaborate effectively with humans (Verma et al., 2023), especially for test-time scaling systems that must constantly decide between refusing to answer and expending further compute to search for a possible solution.

To help address this issue, we evaluate test-time scaling models using a simple class of selection

[1]Code released at https://github.com/wjurayj/final_answer

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 636–644

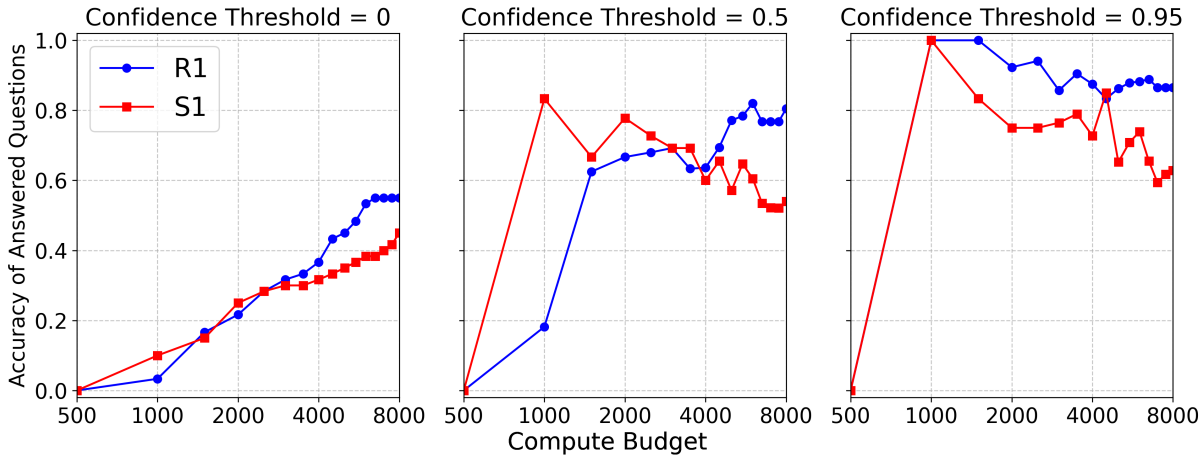


Figure 2: **Confidence thresholds on test-time scaling.** *(left)* When the confidence threshold is 0, the model answers 100% of questions. This is the only performance curve that is reported by test-time scaling research. *(center)* At a moderate threshold, more frequent absentions allow higher response accuracy. *(right)* At a high threshold, small amounts of test-time compute deliver very high accuracy at low answer rates, while test-time scaling provides more answers at the cost of answer accuracy.

functions that reject questions if a model is not confident in its answer after expending its compute budget. We evaluate these systems at different compute budgets, showing a new axis of model performance that answer accuracy alone struggles to measure. We suggest a class of utility functions that represent various levels of error risk to empirically measure the performance of these systems in settings where incorrect answers are penalized. Evaluation in these settings shows how compute scaling affects confidence in existing systems. Based on these insights, we propose a standard method for measuring model performance in settings with non-zero response risk. In summary we:

- Conduct the first evaluation of LLM test-time compute scaling on selective question answering, finding that increasing inference compute can help models distinguish between their correct and incorrect answers. (Section 3)
- Introduce evaluation settings that penalize incorrect answers and allow abstentions to help holistically evaluate models capable of scaling test-time compute. (Section 4)
- Invite the community to report test-time scaling performance on selective question answering under "Jeopardy Odds", which incentivize confidence calibration by penalizing incorrect answers while rewarding correct answers.

2 Methods

We explore how increasing compute budgets affects a model's performance on question answering tasks at different confidence thresholds. The choice of a budget and threshold is a *test-time* decision. We describe methods to quantify the two factors below:

Compute Budget refers to the amount of compute expended by the model at inference time. In all cases, we quantify a model's budget by counting the *number of tokens* in its reasoning trace. We use methods proposed by Muennighoff et al. (2025) to strictly enforce compute budgets. Specifically, we ignore any predicted end-of-thinking delimiters and instead append the token "Wait" if a model attempts to end its reasoning trace before reaching the budget, and we force decode the end-of-thinking delimiter once the budget is reached.

Confidence Threshold refers to the uncertainty of the model in its decoded answer. We quantify a model's confidence as the *sum of the log-probabilities* corresponding to the answer tokens.[2] For a confidence threshold, our selection function (Geifman and El-Yaniv, 2017) only accepts answers that the model delivers with confidence greater than its threshold, abstaining otherwise.

3 Experiments

3.1 Experimental Setup

We evaluate Deepseek-R1-32B (DeepSeek-AI et al., 2025) and s1 (Muennighoff et al., 2025) due to their exhibited test-time scaling capabilities and open-weight checkpoints, and choose

[2]Every answer in our dataset is a 3-digit number between 000 and 999, so consists of the same number of tokens.

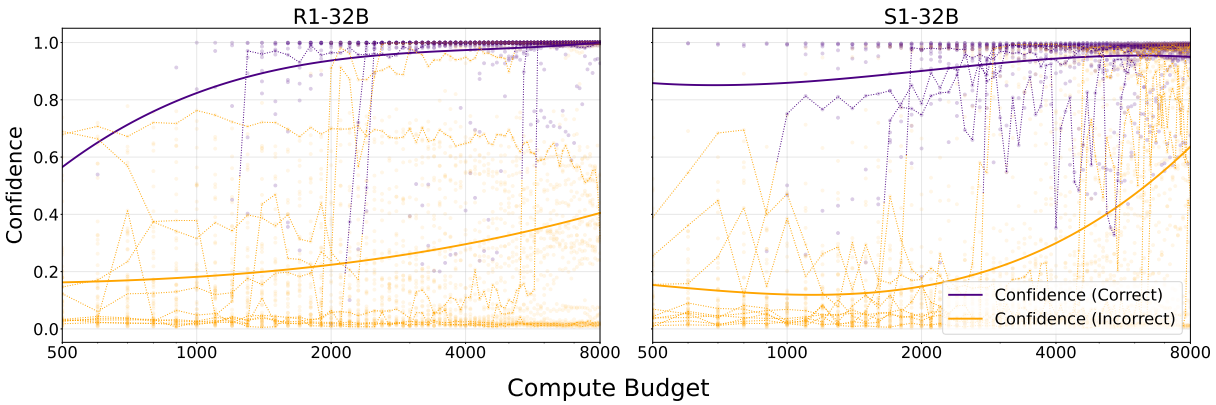


Figure 3: **Test-time scaling improves confidence in correct answers on R1-32B (left) and S1-32B (right).** Each dot represents the model's confidence in an answer after spending a fixed amount of compute. Indigo series are correct answers, while orange series are incorrect. Dotted lines plot confidence trajectories for 10 randomly selected questions, emphasizing how confidence in correct versus incorrect answers changes with test-time compute. Note that individual answers may turn from orange to indigo if the model changes its prediction after thinking longer.

AIME 2024 and 2025 as our primary evaluation dataset. This dataset contains 60 hard math problems on which performance substantially benefits from larger compute budgets, making it a popular benchmark for evaluating test-time scaling. Additional experiments on GPQA (Rein et al., 2024) are included in Appendix B. We test the set of confidence thresholds $\{0.0, 0.5, 0.95\}$ across compute budgets within the range $[500, 8000]$, incrementing by 100 tokens. For a given budget and threshold, we report the accuracy of *answered* questions, treating never answering as yielding accuracy 0. As the number of answered questions differs across confidence thresholds, we note that accuracies are *not directly comparable* between models and compute budgets.

We use widely available open-source libraries to run our experiments, including HuggingFace Transformers (Wolf et al., 2020) and vLLM (Kwon et al., 2023) for language model inference, and the Language Model Evaluation Harness (Gao et al., 2024) to sample reasoning chains at temperature 0 and 32-bit precision. In particular, we use the variant of this library released by Muennighoff et al. (2025), and run a subset of the experiments that they run. We run experiments on 4 H100 GPUs.

3.2 Results

Figure 2 compares the accuracy of answers provided by R1-32B and S1-32B at different test-time compute budgets. When the confidence threshold is 0, models answer every question, so accuracy increases consistently with compute budget. We observe that these subplots are slices of a surface parameterized by compute budget and confidence threshold, shown in Figure 1. While higher confidence thresholds prevent the model from answering at low budgets, scaling compute at high thresholds delivers a larger volume of accurate answers. However, at higher confidence thresholds increased compute budget can actually decrease answer accuracy. This decrease in accuracy of yielded answers does not necessarily reflect decreased performance at higher budgets, but instead that the additional questions answered are less likely to be correct than those answered at lower budgets.

To investigate whether excessive thinking harms accuracy drops by pushing models to abandon correct answers, we plot how a model's confidence in individual answers moves over time. Figure 3 shows the answer confidences given by both models at varying compute budgets, colored according to their correctness, with a curve fit to the distribution. We note that as compute budget increases, the average confidence of its correct answers increases even as additional correct answers are discovered. Notably, this is not a universal property of test-time scaling models: S1-32B does not separate its correct answers from its incorrect answers as well as R1-32B.

4 Utility

4.1 Motivation

When refusal to answer is an option, accuracy can be trivially optimized by a system that answers extremely infrequently. Thus, a useful metric must capture both the accuracy of answers provided and

the system's propensity to provide answers. Many real world scenarios reward correct answers, but incur measurable costs for incorrect answers. We show our results involving confidence thresholds can be adapted to these settings.

Given a model $\mathcal{M}$ and an instance x of a task t, we define a *utility function* f to be

$$f(\mathcal{M}, x) = \begin{cases} 1 & \mathcal{M} \text{ answers } x \text{ correctly} \\ 0 & \mathcal{M} \text{ abstains from answering } x \\ r_t & \mathcal{M} \text{ answers } x \text{ incorrectly} \end{cases}$$

We can assume the reward for correct answers is 1 without loss of generality due to scaling. While there exist scenarios where refusing to answer also incurs a cost, this paper will only discuss the consequences when no extra cost is incurred; the conclusions we draw can be extended to these cases.

4.2 Problem Scenarios

We discuss three settings with varying risk levels:

- Exam Odds ($r_t = 0$): There are no costs incurred by incorrect answers. These are tasks where guessing isn't punished and the model should always try to provide a solution.
- Jeopardy Odds[3] ($r_t = -1$): The cost of an incorrect answer is equal to the reward for a correct answer. In these scenarios, no answer at all is preferable to an incorrect answer.
- High-Stakes Odds ($r_t = -20$): The cost of an incorrect answer far outweighs the reward for a correct answer. In this case, the model should answer only if absolutely certain.

4.3 Results

We keep the same experimental setup as described in Section 3.1. Rather than reporting accuracy, we instead report the utility in the three scenarios above, shown in Figure 4. We focus on the Jeopardy setting because it highlights why a system might choose not to answer; results in the other settings are in Appendix A.

The Exam setting's utility function does not distinguish refusal from incorrectness, so optimal performance is achieved trivially at confidence threshold to 0 so that every question gets the model's best guess. In the Jeopardy setting, however, this is non-trivial. We illustrate the complete function mapping compute budget and confidence threshold to Jeopardy performance in Figure 4: the checkered lines

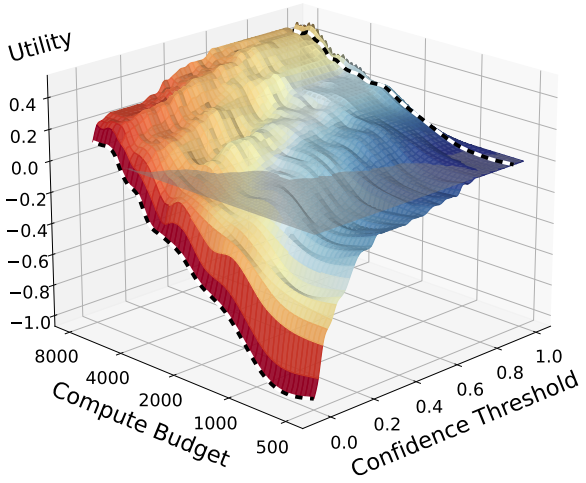


Figure 4: **Utility Surface of DeepSeek R1-32B for Jeopardy.** The vertical axis indicates performance in the Jeopardy setting at different compute budgets and confidence thresholds. The color indicates the proportion of questions that are answered, as in Figure 1. The horizontal plane divides positive and negative utility regions of the operating curve. The checkered lines show the confidence slices that we compare to s1 in Figure 5.

on this surface indicate the two slices that compose R1-32B's portion of Figure 5. We do not suggest that our choice of 0.95 is the optimal threshold for this task, or even that a threshold is the right approach to confidence calibration. Rather, we apply this naive method to show how test-time scaling for selective classification can benefit a practical question-answering setting.

We see on the left of Figure 2 that in the commonly reported Exam Odds, R1-32B and S1-32B scale comparably at threshold 0. In Jeopardy Odds, selective question answering at threshold 0.95 dramatically improves performance for both models. Additionally, although the two models scale comparably at Exam Odds, R1-32B substantially outperforms S1-32B at larger budgets in this new evaluation setting. Previous work overlooks this comparison. We call on future test-time compute scaling research to report optimal utility at Jeopardy Odds in addition to Exam Odds, to help readers understand performance across confidence demands.

5 Related Work

As scaling training compute has become prohibitively expensive (Hoffmann et al., 2022), models that scale performance with test-time compute have become a new frontier (Snell et al., 2024; Wu et al., 2025). These methods have delivered state-of-the-art results on hard reasoning tasks us-

[3]Inspired by the wagers made in the game show's 'Final Jeopardy' stage

ing lengthy chains of thought (DeepSeek-AI et al., 2025; Muennighoff et al., 2025). Current work in this space optimizes for question answering tasks which do not penalize incorrectness, ignoring settings that favor refusal over wrong answers (Ferrucci et al., 2010; Rajpurkar et al., 2018; Kamath et al., 2020). We draw motivation from methods for cost-sensitive learning (Mienye and Sun, 2021) and selective classification (Geifman and El-Yaniv, 2017), which navigate penalties for failure. These settings reward confidence calibration, which can be critical for effective collaboration with human experts (Verma et al., 2023). We are the first to investigate how serialized test-time compute helps models identify when they should not answer.

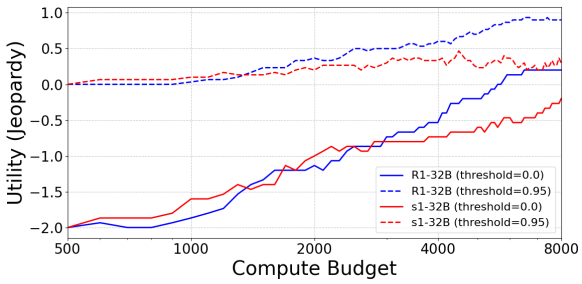


Figure 5: **Jeopardy utility scales differently across models and thresholds.** Performance of S1-32B and R1-32B in the Jeopardy odds setting under different confidence thresholds. Although S1 is competitive at lower budgets when the confidence threshold is 0, a higher threshold shows R1's superior scaling performance in a selective setting.

5.1 Test-time Scaling

Many methods for scaling test-time compute have been explored. These include searching over possible generations (Wang et al., 2024), sampling many completions and selecting the best answer among them (Wu et al., 2025), making gradient updates at inference time (Akyürek et al., 2024; Li et al., 2025), using reinforcement learning to incentivize generating chains of thought before answering (DeepSeek-AI et al., 2025), and simply fine-tuning on longer chains of thought (Muennighoff et al., 2025). Our work considers models fine-tuned on very long reasoning chains, and augments them with the ability to refuse to answer questions where they lack confidence. Concurrent work finds that such models can learn to predict when their answers are unlikely to be correct (Zhang et al., 2025; Huang et al., 2025), but do not show how this affects performance as additional compute is expended. In contrast, we show how model confidence scales with test-time compute, and demonstrate its value for optimizing performance in settings that allow refusal to answer.

5.2 Selective Question Answering

Refusing to answer is an important option in many prior works on question answering. SQuAD 2.0 included this feature by asking questions which have no answer, although they treat abstaining from answering as a correct answer in these unanswerable cases (Rajpurkar et al., 2018). Gameshow based research efforts use an approach more closely aligned with ours, which penalizes systems for answering incorrectly to encourage abstentions when a system cannot develop sufficiently high confidence (Ferrucci et al., 2010; Ferrucci, 2012; Boyd-Graber and Börschinger, 2020; Rodriguez et al., 2021). Related to quiz game settings is research into selective classification, which evaluates models performance across the coverage-accuracy curve, rather than at single point (Geifman and El-Yaniv, 2017). These approaches can be useful for avoiding costly errors in high-pressure domains (Khan et al., 2018), under distribution shift (Ren et al., 2023), or when designing systems that defer to expert humans when it lacks confidence that its input will be helpful (Mozannar and Sontag, 2020). Recent research in language modeling has investigated training language models to refuse to answer (Cao, 2024), and this capacity for refusal has become a point of competition among top industrial labs (Wei et al., 2024). However, this line of work does not investigate this behavior in sequential test-time scaling models on reasoning intensive tasks, where a model might find a confident answer given higher compute budgets.

6 Conclusion

We highlight a region of performance that is currently unexplored by test-time scaling research. We encourage the test-time scaling community to adopt these insights by reporting model scaling performance on benchmarks at both Exam Odds and Jeopardy Odds, to highlight their systems ability to scale confidence with test-time compute. Future work should focus on efficiently allocating test-time compute to meet confidence demands, and could investigate how test-time confidence scaling models should decide between extending reasoning and deferring to human experts.

Limitations

The selection function we implement is based entirely on the likelihood that a large language model assigns a series of tokens after thinking, which is not necessarily the optimal method for model confidence estimation. Furthermore, the Chain-of-Thought scaling method we apply may struggle to generalize to problem types that a model has note The method we use for 'budget forcing' (Muennighoff et al., 2025) may diminish performance by abruptly truncating chains of thought and driving the model outside of its training distribution. Concurrent work has introduced more elegant forms of compute budget control (Aggarwal and Welleck, 2025; Hou et al., 2025). Furthermore, we do not consider how compute costs might be incorporated in the model's utility function, which could encourage increased energy consumption. Finally, we recognize that by evaluating only on English questions and answers, we may miss model capabilities or weaknesses in lower-resource languages or in multilingual settings.

Acknowledgments

This work was funded in part by the U.S. National Science Foundation under grant No. 2204926, and by the Defense Advanced Research Project Agency (DARPA) SciFy program (contract No. HR001125C0304) and CODORD program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or DARPA. We thank Miriam Wanner, Zhengping Jiang, Orion Weller, Marc Marone, Alex Martin, and Beepul Bharti for helpful conversations throughout this project.

References

Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. *arXiv preprint*. ArXiv:2503.04697 [cs].

Ekin Akyürek, Mehul Damani, Linlu Qiu, Han Guo, Yoon Kim, and Jacob Andreas. 2024. The Surprising Effectiveness of Test-Time Training for Abstract Reasoning. *arXiv preprint*. ArXiv:2411.07279 [cs].

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online. Association for Computational Linguistics.

Lang Cao. 2024. Learn to Refuse: Making Large Language Models More Controllable and Reliable through Knowledge Scope Limitation and Refusal Mechanism. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3628–3646, Miami, Florida, USA. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

D. A. Ferrucci. 2012. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3.4):1:1–1:15.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. 2010. Building watson: An overview of the deepqa project. *AI Magazine*, 31(3):59–79.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4885–4894, Red Hook, NY, USA. Curran Associates Inc.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. 2025. ThinkPrune: Pruning Long Chain-of-Thought of LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2504.01296 [cs].

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. 2025. Efficient test-time scaling via self-calibration. *Preprint*, arXiv:2503.00031.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A. Sohel, and Roberto Togneri. 2018. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587.

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

Wen-Ding Li, Keya Hu, Carter Larsen, Yuqing Wu, Simon Alford, Caleb Woo, Spencer M. Dunn, Hao Tang, Wei-Long Zheng, Yewen Pu, and Kevin Ellis. 2025. Combining induction and transduction for abstract reasoning. In *The Thirteenth International Conference on Learning Representations*.

Ibomoiye Domor Mienye and Yanxia Sun. 2021. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, 25:100690.

Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint*. ArXiv:2501.19393 [cs].

Northpointe. 2017. Practitioner's guide to compas core.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2021. Quizbowl: The Case for Incremental Question Answering. *arXiv preprint*. ArXiv:1904.04792 [cs].

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Rajeev Verma, Daniel Barrejón, and Eric Nalisnick. 2023. Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles. *arXiv preprint*. ArXiv:2210.16955 [stat].

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah Goodman. 2024. Hypothesis search: Inductive reasoning with language models. In *The Twelfth International Conference on Learning Representations*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *The Thirteenth International Conference on Learning Representations*.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025. Reasoning models know when they're right: Probing hidden states for self-verification. *Preprint*, arXiv:2504.05419.

A Appendix

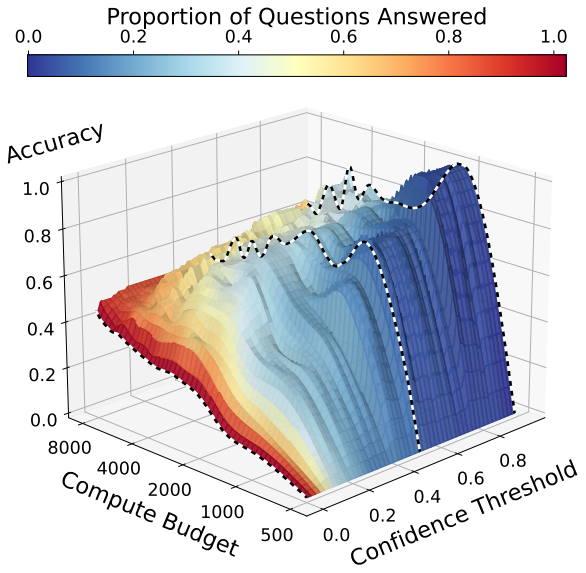
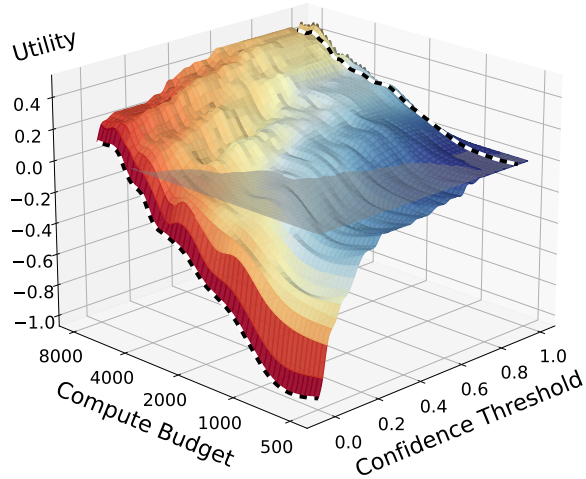


Figure 6: **(above) S1-32B's answer accuracy is a function of compute budget and confidence threshold.** This plot corresponds to the R1-32B plot in Figure 1. **(below) Utility surface of S1-32B for Jeopardy.** This plot corresponds to the R1-32B plot in Figure 4.



B GPQA Experiments

GPQA consists of 448 graduate level multiple choice questions in the fields of physics, biology, and chemistry. These questions are considered "Google-proof", in that skilled non-experts with access to the open internet struggle to answer them. We run experiments on the 'Diamond' subset of this dataset, which consists of the 198 questions which had the clearest answers to domain experts, while being the most difficult for non-experts to answer; this subset has also served as a benchmark for test-time scaling language models (Muennighoff et al., 2025). Our experiments on GPQA follow

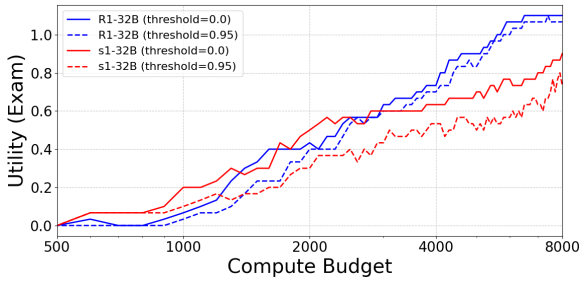
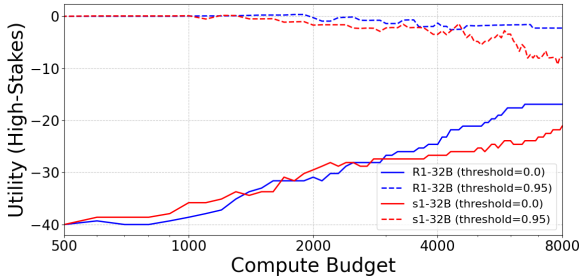


Figure 7: **Additional Model Comparisons.** We additionally compare performance of S1-32B and R1-32B in the Exam Odds *(above)* and High-Stakes Odds *(below)* settings under different confidence thresholds. Like Jeopardy odds depicted in Figure 5, High-Stakes Odds illustrates a performance distinction at high confidence thresholds that is not evident from conventional Exam odds.



the same basic procedure described in Section 3, except that we stop evaluate token budgets in range [500, 4000].

As with AIME, Figure 8 shows how models can increase performance by only answering when highly confident. Notably, there is very little gap between thresholds 0 and 0.5 on GPQA, owing to its multiple choice format. Moreover, whereas R1-32B outperformed S1-32B at all confidence thresholds, R1-32B only exceeds S1-32B at a confidence threshold of 0.95.

Figure 9 shows performance at Jeopardy odds of R1-32B (above) and S1-32B (below). Although both models perform similarly at low thresholds, R1-32B's superior test-time scaling behavior becomes evident at higher confidence thresholds, allowing it to achieve higher peak utilities compared to S1-32B. This gap is reflected in the slices shown in Figure 10. We also note that the slices of the surface are nearly identical at lower confidence thresholds below 0.25. This is likely due to the multiple choice format of GPQA; when the model does not know the answer to a question, it assigns roughly equal probability to the four possible answers.

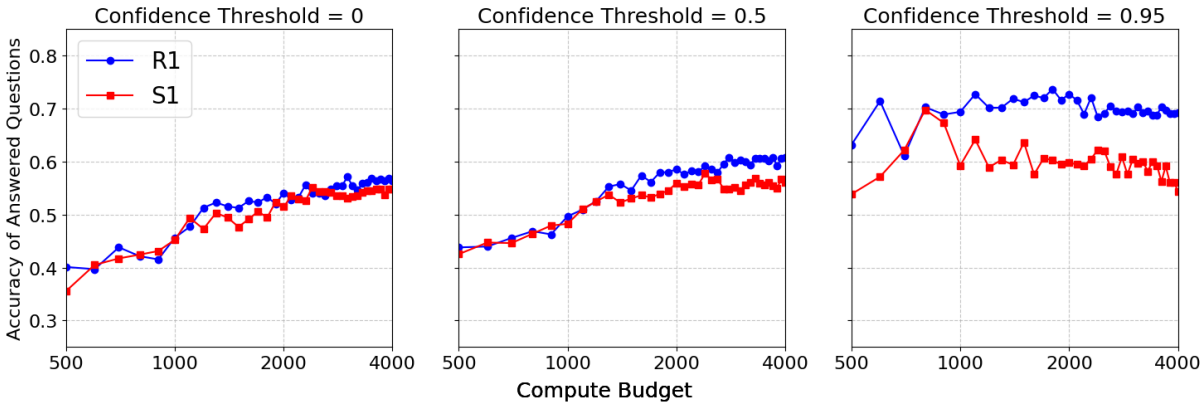


Figure 8: **Confidence thresholds on test-time scaling.** *(left)* When the confidence threshold is 0, the model answers 100% of questions. This is the only performance curve that is reported by test-time scaling research. *(center)* At a moderate threshold, more frequent absentions allow higher response accuracy. *(right)* At a high threshold, small amounts of test-time compute deliver very high accuracy, while test-time scaling provides more answers at the cost of answer accuracy. We treat the decision to never answer as yielding accuracy 0.

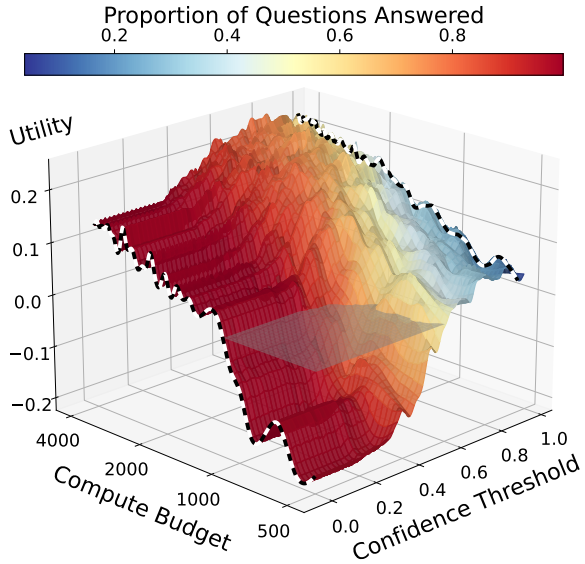
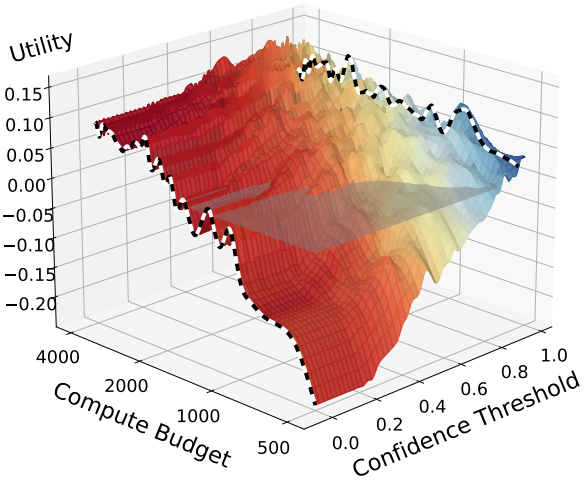


Figure 9: **Utility surfaces of R1-32B (above) and S1-32B (below) for Jeopardy utility on GPQA**. Utility is a function of compute budget and confidence threshold. These plots mirror the surfaces in Figure 4 and Figure 6.



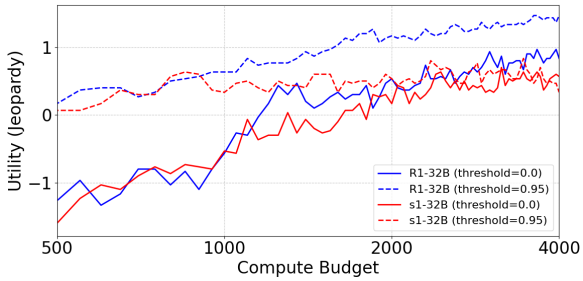


Figure 10: **Jeopardy utility on GPQA scales differently across models and thresholds.** Performance of S1-32B and R1-32B on GPQA in the Jeopardy odds setting under different confidence thresholds. While S1 is competitive in the case when threshold is 0, a higher threshold shows R1's superior scaling performance.