Scaling Agreement with Completions: A Diagnostic Evaluation of LLM Consistency under Inference-Time Budget Constraints

Daivik Patel, Shrenik Patel

Abstract

We investigate the internal consistency of a large language model (Llama3.2) under varying inference-time compute budgets. Building on recent work such as *Thinking Slow, Fast* (Dao et al., 2024), which shows the benefits of generating multiple completions per prompt, we explore not only final accuracy but also internal agreement among completions. Using a subset of 10 math word problems from GSM8K, we vary the number of sampled completions (k = 1, 3, 5), extract final answers, and analyze agreement ratios and inference times. Our results show that consistency increases with k, but at a significant computational cost. We contribute a new diagnostic perspective, agreement analysis, that complements traditional majority voting or pass@k metrics, and provide evidence that such diagnostics reveal deeper properties of LLM behavior under budget constraints.

1 Introduction

Recent advancements in inference-time scaling for large language models (LLMs) have focused on boosting reasoning performance by generating and aggregating multiple completions using methods such as Chain-of-Thought (CoT) reasoning. The ”Thinking Slow, Fast” framework [3] demonstrates that smaller, faster models can outperform larger models under fixed compute budgets by leveraging throughput. However, such work largely evaluates models using coverage and majority-voting accuracy. Here, we investigate a complementary but underexplored question: how consistent is a model internally when producing multiple outputs?

This work aims to introduce and evaluate *agreement* as a diagnostic tool defined as the proportion of completions that agree on the final answer. We explore whether agreement increases with more completions, how it relates to accuracy, and what it reveals about problem difficulty and model stability.

2 Related Work

Inference-time compute scaling has become a powerful method to improve LLM reasoning [1, 4]. Traditional evaluation uses pass@k and majority voting, but these metrics mask internal divergence across samples. Diagnostic metrics such as confidence calibration [2] or process reward models [5] offer finer insight but require additional models or labels. We aim to provide an intrinsic diagnostic method requiring no labels: *agreement among self-generated answers.*

3 Experimental Design

3.1 Model and Dataset

We use the Llama3.2 model with standard temperature sampling (temperature = 1.0, max_new_tokens = 1000) to generate completions for 10 hand-picked GSM8K math word problems. Each problem includes a ground truth answer for reference.

3.2 Generation Strategy

For each problem, we sample $k \in \{1, 3, 5\}$ completions independently using the same prompt. We parse the final answer from each output and compute:

- **Final Answer**: the model's chosen answer after aggregation (majority vote)
- **Agreement Ratio**: proportion of completions that agree with the majority
- **Inference Time**: wall-clock time required to generate all k completions

3.3 Metrics

We report:

- Mean agreement ratio over all 10 problems
- Number of problems with unanimous agreement
- Distribution of agreement counts per problem
- Time vs. Agreement curve

4 Results

4.1 Agreement Improves with k

Agreement improves from 0.30 at $k = 1$ to 0.40 at $k = 5$, suggesting that sampling multiple completions stabilizes the model's output. However, the gain is modest compared to the increase in inference time.

4.2 Inference Time Scaling

As shown in Figure 1, inference time scales approximately linearly with k. Time increases from ~ 46 seconds at $k = 1$ to ~ 240 seconds at $k = 5$.

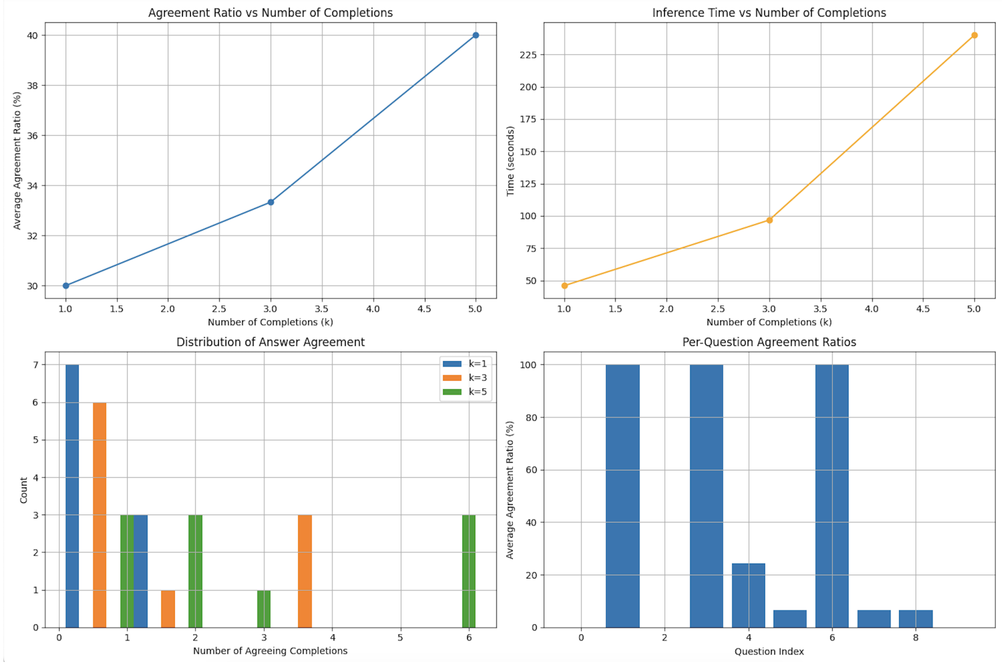

Figure 1: Agreement and inference time analysis for Llama3.2 on 10 GSM8K-style problems across different numbers of completions (k).

4.3 Per-Problem Agreement Distribution

We observe high variance in agreement per question. Some problems elicit consistent completions even at $k = 1$, while others remain unstable even at $k = 5$, indicating differing sensitivity to prompt or inherent ambiguity in reasoning.

5 Discussion

5.1 Agreement as a Diagnostic Tool

Agreement is straightforward to compute, requires no external labels or supervision, and can be applied across tasks and models. Unlike traditional evaluation metrics that rely on ground truth or external verifiers, agreement leverages the model's own outputs to reflect its internal uncertainty. A high agreement ratio across completions suggests a degree of output determinism or confidence in reasoning, while low agreement highlights internal variability, model confusion, or sensitivity to prompt phrasing.

Furthermore, analyzing per-question agreement distributions allows practitioners to isolate failure modes in a granular way. For example, even if the final prediction is incorrect, consistent agreement across completions may suggest that the model has internalized a stable (though flawed) reasoning path.

5.2 Complementing Existing Metrics

Most current evaluations focus on whether the final answer is correct (accuracy) or if it appears within a set of completions (coverage/pass@k). These metrics, while useful, collapse multiple generations into a binary outcome. In contrast, agreement provides insight into how *confidently* or *unanimously* the model reaches that decision.

High accuracy with low agreement could suggest a correct answer was reached by chance among diverse generations. Conversely, high agreement but low accuracy suggests a consistent, but incorrect, reasoning pattern. This can be useful for identifying systemic errors or biases in the model's training data or inductive biases or in applications where correctness is difficult to judge (e.g., creative writing, ambiguous queries).

5.3 Limitations and Future Work

This study is constrained by a small dataset (10 GSM8K problems) and a single model variant (Llama3.2). While the observed trends are promising, generalization requires broader experimentation. Larger and more diverse datasets, such as full GSM8K or MATH, should be used to validate the robustness of agreement-based insights.

Additional model architectures (e.g., Mamba hybrids, distilled Transformers) should also be tested to examine whether agreement behavior is architecture-dependent. Furthermore, exploring the correlation between agreement and other factors such as CoT length, perplexity, or entropy, could uncover deeper connections between model behavior and reasoning stability.

Future work could also explore richer agreement measures beyond final answer comparison. For example, agreement over intermediate reasoning steps, logical structure, or latent representations may reveal even more about model confidence and error types. Combining agreement with reward models or verifier scores could lead to new hybrid aggregation strategies that balance correctness and consistency.

6 Conclusion

We introduced agreement ratio as a diagnostic metric to measure internal consistency in LLM completions under inference-time scaling. Our experiments with Llama3.2 on GSM8K-style problems show that agreement increases with more completions but at rising computational cost. We provide a new angle on evaluating LLMs that complements standard accuracy and opens new research directions in reasoning reliability.

Building on the *Thinking Slow, Fast* framework [3], which focuses on coverage and accuracy under compute constraints, we contribute a finer-grained introspection into how consistent model outputs are across completions. While Tri Dao et al. demonstrate throughput-driven performance improvements, our work reveals what happens within those completions, offering a new lens for understanding model behavior and failure modes even in small-scale deployments. This agreement-based view enhances the interpretability and diagnostic power of test-time compute scaling techniques.

Acknowledgements

Thanks to Tri Dao et al. for the original *Thinking Slow, Fast* framework that inspired this work.

References

[1] Ben Brown, Jonah Juravsky, Rachel Ehrlich, Rylan Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

[2] Stephanie Lin, Jacob Hilton, and Amanda Askell. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022.

[3] Daniele Paliotta, Junxiong Wang, Matteo Pagliardini, Kevin Y Li, Aviv Bick, J Zico Kolter, Albert Gu, François Fleuret, and Tri Dao. Thinking slow, fast: Scaling inference compute with distilled reasoners. *arXiv preprint arXiv:2502.20339*, 2025.

[4] Andrew Snell et al. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2401.00001*, 2024.

[5] Jonathan Uesato, Xuezhi Jiang, Aitor Lewkowycz, Jiecao Chen, Eli Zelikman, Stanislaw Jastrzebski, Catherine Olsson, et al. Solving math word problems with process supervision. In *International Conference on Learning Representations (ICLR)*, 2022.