# Robustness Testing and Comparing Reasoning Techniques for Small Language Models

Study Project Exposé

Leon Wagner
Institute for Computer Science, Humboldt-Universität zu Berlin
wagnerql@hu-berlin.de

November 13, 2024

## 1 Introduction

Artificial Intelligence today has advanced to a point where it can almost effectively grasp and understand natural human language [1], opening a wide range of new opportunities across fields such as communication, automation, coding, information retrieval, and personalized user experiences. Large Language Models (LLMs) are trained on extensive datasets representing a broad spectrum of human knowledge up to a specific timestamp. The most advanced models, such as ChatGPT-4o or Claude 3.5 Sonnet, can often outperform untrained humans in analyzing and processing information across diverse domains. However, their performance remains limited by the scope of their training data and an inherent lack of real-world experience [2], particularly in fields requiring rigorous logical reasoning [3], such as mathematics and complex problem-solving.

These capabilities have been achieved primarily through advancements in architecture, such as the Transformer model and self-attention mechanisms, as well as through massive scaling of training data, reinforcement learning from human feedback, and improvements in reasoning abilities [4]. These reasoning abilities are becoming increasingly important, as further scaling of model size yields diminishing returns in improving the model's inference capabilities [5] [6]. Broadly defined, reasoning is the process of analyzing something logically and methodically, drawing on evidence and previous experiences to arrive at a conclusion or make a decision [7]. While the concept of reasoning in language models is not new, it still lacks a precise definition, and "reasoning" often refers to informal reasoning in research contexts, though this distinction is not always specified [8].

As reasoning approaches become more significant, they offer potential use cases for achieving comparable inference results in much smaller language models (SLMs) [9] containing only 1–7 billion parameters. Furthermore the continued scaling of LLMs has made them increasingly challenging for researchers to replicate and experiment with at universities, as academic institutions often lack the financial and computational resources of major tech companies like Google, Meta, or Microsoft. To address this, researchers

at Microsoft experimented with reasoning methods on SLMs and claim that their SLM, when enhanced with reasoning techniques, achieved inference results comparable to those of models 10 times larger that did not use reasoning [10].

**Problem Statement:** Despite promising results from reasoning-enhanced models, uncertainties remain regarding whether these significant performance improvements are as substantial as their claims. Do these performances solely stem from reasoning methods or do other underlying factors contribute [11]. Additionally, it is uncertain whether these reasoning approaches can achieve similar efficacy when applied to language models with as few as 1 billion parameters or smaller [12] [13]. There is also limited understanding of the full range of reasoning techniques applicable to SLMs and the extent to which these methods can be effectively combined [14] [15]. This project aims to address these gaps by examining the potential and limitations of reasoning approaches specifically tailored for SLMs.

# 2 Research Goal

The aim is to develop, pen-test, and compare three reasoning models, each leveraging one of the following distinct approaches to reasoning (based on distinctions proposed by Huang et al. [8]) on a 1 and 3-billion parameter language model:

1. **Multi-Stage Finetuning:** In the Orca 2 paper [10], the model enhances reasoning in SLMs through a specialized finetuning approach that incorporates multiple reasoning techniques. The model is trained on a diverse dataset featuring methods like step-by-step processing, recall-then-generate, and cautious reasoning, extracted from larger models (e.g., GPT-4). During finetuning, Orca 2 uses a technique called "prompt erasure," where specific task instructions are removed, encouraging the model to learn reasoning strategies independently rather than by direct imitation. In this study project the multi-stage fine-tuning process with the FLAN-v2 dataset and prompt erasure of the Orca 2 paper will be replicated on the Llama 3.2 model. Since the paper has very promising results in enhancing reasoning capabilities in SLMs, enabling them to rival much larger models in complex tasks, there is a high interest to further investigate and verify the benefits from their reasoning method.

2. **Prompting and In-Context Learning:** Chain-of-Thought (CoT) prompting and Self-Consistency became two of the most famous methods for enhancing reasoning in language models that work just by prompting and answer aggregation. CoT prompting encourages the model to generate intermediate steps, mimicking human-like reasoning, which helps in solving multi-step tasks by breaking them down into smaller, manageable parts. Self-consistency builds on this by sampling multiple reasoning paths for the same prompt and selecting the most consistent answer through a majority vote (good for reducing hallucinations of models). Applying this to a smaller model like the 1 and 3-billion parameter Llama 3.2 involves generating diverse reasoning paths, aggregating results, and choosing the answer with the highest agreement across paths. This approach became very popular because it aims to enhance reasoning without the need for any additional training. Key references for this approach include *Self-Consistency Improves Chain of Thought Reasoning in Language Models* [16] and the knowledge distillation approach in *Teaching Small Language Models to Reason* [17]. Additionally, the concept of "Problem Decomposition" from *Least-to-Most Prompting* [18] to decompose a very

complex problem into smaller subproblems (similar to "Divide and Conquer") in order to solve them in specific order also sounds very promising.

3. **Hybrid Method:** The Hybrid Method combines prompting and fine-tuning, with a recent approach gaining interest since 2022: bootstrapping and self-improvement through the *Self-Taught Reasoner (STaR)* method [19]. Rather than traditional fine-tuning on pre-existing datasets, STaR enables a model to self-improve through iterative bootstrapping. The model first generates detailed rationales, or step-by-step explanations, to clarify its thought process for each answer. It is then fine-tuned on its own rationales that led to correct answers, repeating this process iteratively to enhance model performance. Each cycle produces better rationales and, in turn, improves training data quality, yielding an overall more capable model. Given the uniqueness and promise of this approach, evaluating its effectiveness on the smaller Llama 3.2 model and comparing it with other methods presents a valuable research opportunity.

| **Method** | **Advantages** | **Disadvantages** |
|---|---|---|
| Multi-Stage Finetuning | - Combines various reasoning techniques<br>- Prompt Erasure fosters independent strategies<br>- High performance, even in complex tasks | - Resource-intensive data collection and fine-tuning<br>- Relies on larger models for initial strategies |
| Prompting and In-Context-Learning | - Efficient, needs no additional training or data<br>- Self-Consistency improves accuracy and reduces hallucinations<br>- Adaptable for small models with distilled knowledge | - Performance heavily depends on model size<br>- Inconsistent accuracy on complex tasks<br>- don't improve reasoning capabilities themselves, since parameters of model remain unchanged |
| Hybrid Method | - Self-improvement through iterative learning<br>- Detailed reasoning enhances answer quality<br>- Comparable results to larger models | - Slow learning due to repeated iterations<br>- Complex setup and resource-intensive |

Table 1: Comparison of Advantages and Disadvantages of Reasoning Approaches

Evaluation metrics will be used to assess and compare the performance and robustness of the models, providing a clear measure of the project's success in meeting its goals. Details are provided in the Evaluation section.

# 3 Approach

## 3.1 Literature Review

A comprehensive review of the reasoning techniques will be conducted, focusing on their applicability and potential adaptations required for integration with the Llama 3.2 model.

### 3.2 Implementation and Optimization

Each reasoning technique will be implemented within a separate instance of the Llama 3.2 model, with modifications noted to document any deviations from the original methodologies. For comparative analysis, both an unmodified Llama 3.2 model (1.2 billion parameters) and a larger variant (3.2 billion parameters) will be included in the project.

## 4 Evaluation

### 4.1 Robustness Testing

To evaluate the resilience and robustness of each reasoning technique, two key metrics are employed:

- **Semantic Consistency Score:** This metric measures how often a language model reaches the same overall conclusion or final answer when presented with slight variations in input [20].
- **Logical Contradiction Rate (LCR):** This metric captures the frequency of logical contradictions within generated reasoning chains. A lower LCR suggests more consistent and logically sound outputs [21].

These metrics provide a clear view of the model's robustness by examining its vulnerability to noisy scenarios, consistency in logical reasoning, and stability under diverse input conditions. This testing phase will guide refinements to the reasoning framework, enhancing robustness and reliability.

### 4.2 Performance Evaluation

In the final step, the performance boost of the Llama 3.2 model with reasoning is compared to the base Llama 3.2 model without reasoning across several evaluation benchmarks like AGIEval, DROP, RACE, ARC, GSM8K and MMLU as well as a comparison to the model with a larger model like Llama 3.2 (3b). Key metrics are collected and analyzed.

## 5 Project Submission

The project will be submitted via a Git repository, including the evaluation metrics, a report, and the three models implementing distinct reasoning methods.

## References

[1] H. Dhamelia, "Unlocking semantic dimensions: Harnessing ai for next-gen natural language understanding," *International Journal for Research in Applied Science and Engineering Technology*, 2023.

[2] R. Shiffrin and M. Mitchell, "Probing the psychology of ai models," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 120, 2023.

[3] A. Kalyan, A. Kumar, A. Chandrasekaran, A. Sabharwal, and P. Clark, "How much coffee was consumed during emnlp 2019? fermi problems: A new reasoning challenge for ai," *ArXiv*, vol. abs/2110.14207, 2021.

[4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[5] V. Udandarao, A. Prabhu, A. Ghosh, Y. Sharma, P. H. S. Torr, A. Bibi, S. Albanie, and M. Bethge, "No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance," 2024. [Online]. Available: https://arxiv.org/abs/2404.04125

[6] E. Caballero, K. Gupta, I. Rish, and D. Krueger, "Broken neural scaling laws," *ArXiv*, vol. abs/2210.14891, 2022.

[7] P. Wason, *Psychology of Reasoning: Structure and Content*. Cambridge/Harvard University Press, 1972.

[8] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2022.

[9] X. Zhao, Y. Xie, K. Kawaguchi, J. He, and Q. Xie, "Automatic model selection with large language models for reasoning," pp. 758–783, 2023.

[10] A. Mitra, L. D. Corro, S. Mahajan, A. Codas, C. Simoes, S. Agarwal, X. Chen, A. Razdaibiedina, E. Jones, K. Aggarwal, H. Palangi, G. Zheng, C. Rosset, H. Khanpour, and A. Awadallah, "Orca 2: Teaching small language models how to reason," 2023. [Online]. Available: https://arxiv.org/abs/2311.11045

[11] Z. Xi, S. Jin, Y. Zhou, R. Zheng, S. Gao, T. Gui, Q. Zhang, and X. Huang, "Self-polish: Enhance reasoning in large language models via problem refinement," pp. 11 383–11 406, 2023.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[13] Z. Yuan, H. Yuan, C. Li, G. Dong, C. Tan, and C. Zhou, "Scaling relationship on learning mathematical reasoning with large language models," *ArXiv*, vol. abs/2308.01825, 2023.

[14] M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang, "Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks," *ArXiv*, vol. abs/2305.18395, 2023.

[15] K. Shridhar, A. Stolfo, and M. Sachan, "Distilling reasoning capabilities into smaller language models," pp. 7059–7073, 2022.

[16] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," 2023. [Online]. Available: https://arxiv.org/abs/2203.11171

[17] L. C. Magister, J. Mallinson, J. Adamek, E. Malmi, and A. Severyn, "Teaching small language models to reason," 2023. [Online]. Available: https://arxiv.org/abs/2212.08410

[18] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, and E. Chi, "Least-to-most prompting enables complex reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2205.10625

[19] E. Zelikman, Y. Wu, and N. D. Goodman, "Star: Self-taught reasoner," *arXiv preprint arXiv:2203.14465*, 2022.

[20] H. Raj, D. Rosati, and S. Majumdar, "Measuring reliability of large language models through semantic consistency," *ArXiv*, vol. abs/2211.05853, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253499179

[21] N. Mündler, J. He, S. Jenko, and M. T. Vechev, "Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation," *ArXiv*, vol. abs/2305.15852, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258887694