Reasoning and Knowledge Representation in Small Language Models: An Interpretability Analysis on the ARC-AGI Benchmark

Daniel Brito dos Santos

May 12, 2025

Introduction

The rapid advancement of Large Language Models (LLMs) has revolutionized numerous fields, yet their immense scale often obscures the fundamental mechanisms underlying their capabilities (Templeton et al. 2024). Currently, small language models (SLM) have emerged as powerful, efficient, and accessible alternatives, offering a unique lens to study the core aspects of AI (DeepSeek-AI et al. 2025; Qwen Team et al. 2025). This research focuses on SLMs, specifically recent models around 1.5 billion parameters (Team 2025), to investigate their emergent reasoning and representation of internal knowledge.

Understanding the internals of the language model is essential for robust, reliable, and trustworthy AI. Interpretability research aims to demystify these "black boxes," explaining *how* and *why* models produce specific results (Doshi-Velez and Kim 2017; Amodei 2025). Compared to their larger counterparts, SLMs provide a more accessible setting for in-depth interpretability research. Their relative simplicity enables more precise analysis of learned representations and the computational 'circuits' underlying model behavior (Elhage et al. 2021; Olah et al. 2020; Nanda, Chan, et al. 2023; Lindsey et al. 2025).

The Abstraction and Reasoning Corpus (ARC-AGI) benchmark (Chollet 2019; ARC Prize 2024) is distinctly advantageous for this pursuit. ARC-AGI tasks demand genuine reasoning and abstraction, probing human-like fluid intelligence beyond learned textual patterns. Its well-defined, visual tasks are ideal for studying how models build and manipulate abstract concepts. Investigating SLM behavior on ARC-AGI can yield significant insights into reasoning primitives, where even advanced LLMs struggle (Xu et al. 2023; Ichter et al. 2023).

This research intersects efficient SLMs, AI interpretability, and robust reasoning via ARC-AGI. Understanding how SLMs tackle ARC-AGI can help develop more capable, explainable AI, fostering safer deployments.

- **General Objective:** To investigate and causally explain specific reasoning failure modes in an SLM (e.g., a Qwen model of 1.5B parameters) on curated ARC-AGI problems, focusing on the interplay between input representation and internal mechanisms.

- **Specific Objectives:**

  1. Develop and evaluate effective textual/symbolic representations for curated ARC-AGI tasks, enabling SLM processing, and analyze their impact on model behavior.

2. Evaluate SLM performance on this curated ARC-AGI subset (selected for tractable structure and specific reasoning patterns like symmetry or counting) to identify consistent, interpretable failure modes.
3. Conduct targeted, in-depth failure analysis on isolated tasks using causal interpretability techniques (primarily activation patching and interventions) to identify why internal mechanisms break down, potentially starting with CoT analysis.
4. Provide robust causal explanations for specific reasoning failures and, where feasible, show how minimal interventions predictably alter model behavior concerning these failures.
5. Contribute insights into representational and mechanistic prerequisites for robust abstract reasoning in SLMs, highlighting common pitfalls and their potential understanding or mitigation.

Theoretical Framework

Small Language Models (SLMs)

The growing interest in small, efficient language models such as Phi-2 (Javaheripi et al. 2023), the Qwen series (Qwen Team et al. 2025), and Gemma (Google 2024) is driven by the need for lower computational costs, faster inference, and easier deployment. Despite their size, some SLMs exhibit surprising emergent capabilities, including reasoning and in-context learning (Brown et al. 2020), though generally to a lesser extent than larger models. Their compact architecture also makes them well-suited for detailed interpretability research. The Qwen3 family, for instance, offers open-source, Transformer-based (Vaswani et al. 2017) models that achieve strong benchmark performance (Team 2025).

Interpretability in Language Models

Increasing AI model complexity heightens the need to understand their decision processes (Amodei 2025). Methods range from local (e.g., LIME (Ribeiro, Singh, and Guestrin 2016), SHAP (Lundberg and Lee 2017)) to global explanations. For Transformers, techniques include attention analysis (Bahdanau, Cho, and Bengio 2015; Vig and Belinkov 2019), probing activations for knowledge (Alain and Bengio 2016; Hewitt and Manning 2019; Fierro et al. 2025; Wendler et al. 2024), causal mediation analysis (Vig, Gehrmann, et al. 2020; Geiger et al. 2021), and mechanistic interpretability (reverse-engineering circuits) (Olah et al. 2020; Elhage et al. 2021; Nanda, Chan, et al. 2023; Olsson et al. 2022; Kantamneni and Tegmark 2025). Applying these to SLMs helps map abstract task features to neural representations.

The ARC-AGI Benchmark

Proposed by François Chollet (Chollet 2019), ARC-AGI (ARC Prize 2024) evaluates human-like general intelligence through visual reasoning puzzles that require pattern inference from examples. Key features include minimal prior knowledge, an emphasis on abstraction and analogy, and resistance to brute-force solutions. ARC-AGI is critical for interpretability, as success in solving these tasks demonstrates genuine generalization, offering valuable insights into the model's reasoning process. Despite recent advances, LLMs still struggle significantly with these tasks (Xu et al. 2023; Ichter et al. 2023), making ARC-AGI a key frontier for AI reasoning.

Methodology

This research adopts a focused approach to understand reasoning breakdowns in an SLM (e.g., a Qwen 1.5B model) on abstract tasks, using careful task selection, robust input representation, and principled application of causal interpretability methods. Depth is prioritized over breadth.

1. **ARC-AGI Task Curation and Representation Development:** A small, representative subset of ARC-AGI tasks (Chollet 2019; ARC Prize 2024) will be selected for testing fundamental reasoning primitives (e.g., symmetry, counting) and amenability to textual/symbolic representation. Various encoding strategies for these visual tasks will be explored and evaluated for expressiveness and impact on initial model interactions (e.g., via probing or attention analysis), a critical step for meaningful input.

2. **Focused Baseline Evaluation and Failure Mode Identification:** The SLM's performance (zero-shot/few-shot, potentially with CoT (Wei et al. 2023)) on curated, encoded ARC-AGI tasks will be evaluated. The goal is not benchmark success, but identifying tasks with intriguing partial successes or, more likely, consistent, analyzable failure modes for deep interpretability.

3. **Mechanistic Interpretability of Failure Modes:** For selected tasks/failures, focused interpretability techniques will be applied. Initial explorations might involve CoT analysis or attention patterns (Vig and Belinkov 2019), before emphasizing causal methods like activation patching (Zhang and Nanda 2024; Dumas et al. 2024; Nanda, Rajamanoharan, et al. 2023), and potentially path patching/causal scrubbing (Nanda, Chan, et al. 2023; Elhage et al. 2021), to trace information flow and identify components or 'circuits' (Olah et al. 2020) failing reasoning steps. Tools include PyTorch, Hugging Face Transformers (Wolf et al. 2020), TransformerLens (Nanda 2022). Caution will avoid spurious correlations, validating claims with interventions.

4. **Causal Explanation and Experimental Intervention:** Based on mechanistic analysis, hypotheses about causal chains of specific reasoning breakdowns will be formulated. For localized failures, targeted interventions (e.g., fixing activations, patching representations) will test if model behavior changes predictably, validating causal hypotheses.

5. **Consolidation, Synthesis, and Dissemination:** Results will be synthesized into a narrative around the identified failure modes. Conclusions will address the interplay of input representation, model architecture, and emergent reasoning (or failures). The Master's thesis will document findings and insights.

Flexibility and Fallback Strategy: Given ARC-AGI's difficulty and interpretability's exploratory nature, if identifying suitable ARC-AGI tasks with dissectable SLM failure modes proves intractable despite representation efforts, the project will pivot to analyzing SLM behavior on ARC-inspired synthetic tasks. These would have ground-truth reasoning steps, allowing continued rigorous investigation under controlled conditions, ensuring core research questions are addressed.

Timeline

This Master's research is planned over a standard 2-year (4-semester) period, aligning with UFMG's typical program structure:

- **Semester 1:** Intensive literature review (Methodology Stage 1); SLM/ARC-AGI tool selection/setup; foundational AI/ML coursework.
- **Semester 2:** Completion of ARC-AGI task curation/representation evaluation and SLM baseline evaluation (Methodology Stages 1 & 2); initial interpretability experiments (begin Stage 3); further relevant coursework (e.g., NLP, Deep Learning, Reinforcement Learning).
- **Semester 3:** In-depth interpretability analysis and reasoning process investigation (Stages 3 & 4); data analysis/synthesis; significant thesis writing progress.
- **Semester 4:** Completion of experimental work/final analysis (Methodology Stage 4 & begin Stage 5); final thesis writing, revision; defense preparation.

Dissertation writing will be ongoing, concentrated in the second year. Coursework will be selected from UFMG's offerings pertinent to AI, interpretability, and research methods.

References

Alain, Guillaume and Yoshua Bengio (2016). "Understanding intermediate layers using linear classifier probes". In: *arXiv preprint arXiv:1610.01644*.

Amodei, Dario (2025). *The Urgency of Interpretability*. Accessed: 2025-05-13.

ARC Prize (2024). *ARC-AGI Benchmark*. https://arcprize.org/arc-agi. Accessed: 2025-05-13.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural machine translation by jointly learning to align and translate". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv: 1409.0473.

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL].

Chollet, François (2019). "On the measure of intelligence". In: *arXiv preprint arXiv:1911.01547*.

DeepSeek-AI et al. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv: 2501.12948 [cs.CL].

Doshi-Velez, Finale and Been Kim (2017). "Towards A Rigorous Science of Interpretable Machine Learning". In: *arXiv preprint arXiv:1702.08608*.

Dumas, Clément et al. (2024). "How do Llamas process multilingual text? A latent exploration through activation patching". In: *ICML 2024 Workshop on Mechanistic Interpretability*.

Elhage, Nelson et al. (2021). "A Mathematical Framework for Transformer Circuits". In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html.

Fierro, Constanza et al. (2025). *How Do Multilingual Language Models Remember Facts?* arXiv: 2410.14387 [cs.CL].

Geiger, Atticus et al. (2021). "Examining Attributions on Textual Data". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1260–1275.

Google (2024). *Gemma: Open Models Based on Gemini Research and Technology*. https://ai.google.dev/gemma.

Hewitt, John and Christopher D. Manning (2019). "A structural probe for finding syntax in word representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138.

Ichter, Brian et al. (2023). “Can Large Language Models Reason and Plan?” In: *arXiv preprint arXiv:2310.09176*.

Javaheripi, Mojan et al. (Dec. 2023). *Phi-2: The surprising power of small language models*. Microsoft Research Blog.

Kantamneni, Subhash and Max Tegmark (2025). *Language Models Use Trigonometry to Do Addition*. arXiv: 2502.00873 [cs.AI].

Lindsey, Jack et al. (2025). “On the Biology of a Large Language Model”. In: *Transformer Circuits Thread*.

Lundberg, Scott M and Su-In Lee (2017). “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems (NIPS)*. Vol. 30.

Nanda, Neel (2022). *TransformerLens (formerly EasyTransformer)*. GitHub repository. https://github.com/neelnanda-io/TransformerLens.

Nanda, Neel, Lawrence Chan, et al. (2023). “Progress on Interpreting Transformers”. In: *arXiv preprint arXiv:2305.01610*.

Nanda, Neel, Senthooran Rajamanoharan, et al. (Dec. 2023). *Fact Finding: Attempting to Reverse-Engineer Factual Recall on the Neuron Level*.

Olah, Chris et al. (2020). “Zoom In: An Introduction to Circuits”. In: *Distill* 5.3. DOI: 10.23915/distill.00024.001.

Olsson, Catherine et al. (2022). “In-context Learning and Induction Heads”. In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Qwen Team et al. (2025). *Qwen2.5 Technical Report*. arXiv: 2412.15115 [cs.CL].

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. arXiv: 1602.04938 [cs.LG].

Team, Qwen (2025). *Qwen3: Think Deeper, Act Faster*. Accessed: 2025-05-13.

Templeton, Adly et al. (2024). “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet”. In: *Transformer Circuits Thread*.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems (NIPS)*. Vol. 30.

Vig, Jesse and Yonatan Belinkov (2019). “Analyzing the Structure of Attention in a Transformer Language Model”. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 63–76.

Vig, Jesse, Sebastian Gehrmann, et al. (2020). “Causal mediation analysis for interpreting neural NLP: The case of gender bias”. In: *arXiv preprint arXiv:2004.12990*.

Wei, Jason et al. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv: 2201.11903 [cs.CL].

Wendler, Chris et al. (2024). *Do Llamas Work in English? On the Latent Language of Multilingual Transformers*. arXiv: 2402.10588 [cs.CL].

Wolf, Thomas et al. (2020). “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45.

Xu, Yushi et al. (2023). “A Comprehensive Study of Large Language Models on Challenging Benchmarks: ARC, HellaSwag, and MMLU”. In: *arXiv preprint arXiv:2307.08031*.

Zhang, Fred and Neel Nanda (2024). *Towards Best Practices of Activation Patching in Language Models: Metrics and Methods*. arXiv: 2309.16042 [cs.LG].