Evaluating Stability and Interchangeability of Large Language Models in Mathematical Reasoning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Chain-of-Thought (CoT) prompting has significantly advanced the reasoning capabilities of large language models (LLMs). While prior work focuses on improving model performance through internal reasoning strategies, little is known about the interchangeability of reasoning across different models. In this work, we explore whether a partially completed reasoning chain from one model can be reliably continued by another model, either within the same model family or across families. We achieve this by assessing the sufficiency of intermediate reasoning traces as transferable scaffolds for logical coherence and final answer accuracy. We interpret this interchangeability as a means of examining inference-time trustworthiness, probing whether reasoning remains both coherent and reliable under model substitution. Using token-level log-probability thresholds to truncate reasoning at early, mid, and late stages from our baseline models, Gemma-3-4B-IT and LLaMA-3.1-70B-Instruct, we conduct continuation experiments with Gemma-3-1B-IT and LLaMA-3.1-8B-Instruct to test intra-family and cross-family behaviors. Our evaluation pipeline leverages truncation thresholds with a Process Reward Model (PRM), providing a reproducible framework for assessing reasoning stability via model interchange. Evaluations with a PRM reveal that hybrid reasoning chains often preserve, and in some cases even improve, final accuracy and logical structure. Our findings point towards interchangeability as an emerging behavioral property of reasoning models, offering insights into new paradigms for reliable modular reasoning in collaborative AI systems.

1 Introduction

Chain of Thought (CoT) prompting emerged as powerful mechanism to improve the reasoning capabilities of large language models (LLMs) by encouraging intermediate structured reasoning steps before arriving at a final answer [Wei et al., 2023]. Previous work has explored how CoTs improve individual model performance even in zero-shot settings [Kojima et al., 2023, Zhang et al., 2022, Jin et al., 2024]. More recently, Hebenstreit et al. [2024] examined the transferability of entire CoT sequences by evaluating whether rationale prompts discovered on one model could generalize reasoning strategies across a range of models and tasks. However, it remains unclear to what extent reasoning trajectories are interchangeable when only partially reused. In light of this, our aim is to answer the central research question: *To what extent can the modular decomposition of complex mathematical reasoning tasks enhance the zero-shot performance and interpretability of Large Language Models, when utilizing a collaborative framework that includes both intra-family and cross-family LLMs?*

In this work, we investigate the process-level interchangeability in language model reasoning by evaluating how well different models can continue the CoT of another's midstream. We begin with

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

full CoT traces generated by a strong base model (e.g., Gemma-3-4B-IT and LLaMA-3.1-70B-Instruct), recording token-level log-probabilities to guide strategic truncation at 25%, 50%, and 75% of the cumulative log-probability, capturing early, mid, and late stages of reasoning based on informativeness. From these truncated points, alternative models (including those from different families or architectures) are tasked with continuing the reasoning process using only truncated intermediate steps as input We then assess not only accuracy, but also the coherence, semantic alignment, and logical consistency of the full reasoning chain, by using a Process Reward Model (PRM) trained to evaluate multi-step mathematical reasoning performance. Ultimately, our aim is to characterize how steady transferability depends on truncation point, model pairing, and reasoning domain, yielding clearer interpretations into the dynamics of CoT continuation success that goes beyond final answer accuracy.

Whereas prior work has explored how CoT prompting improves reasoning within individual models [Wei et al., 2023], whether reasoning can be interchanged across models mid-process remains largely unexamined.

We provide compelling early evidence that such a handoff is often successful within the same model family. We show that a partially completed CoT from a strong model, such as Gemma-3-4B-IT, can often be continued by another model of similar or lesser capacity within the same family. By leveraging log-probability-based truncation and PRM-based scoring, we found that these hybrid trajectories maintain high coherence and correctness with minimal loss in reasoning quality.

We found that this practice may not be suitable for all cross-family continuation pairings, as some unreliably preserve quality and coherence of the reasoning chain in our experimentation. Our findings expose distinctions across different model architectures and introduce a promising new paradigm for collaborative reasoning, where high-capacity models can be reserved for the most uncertain portions of a problem, allowing lighter models to reliably finish the remainder of the task.

2 Related Works

LLMs generate responses by autoregressively predicting outputs based on the preceding context, which is learned during pre-training [OpenAI et al., 2024]. As a result, their output can fluctuate even when prompted with identical inputs, introducing variability in reasoning trajectories [Amatriain, 2024]. This, coupled with the absence of structured reasoning mechanisms, often leads to inconsistencies in multistep logical inference. Consequently, assessing the reliability and soundness of their reasoning becomes increasingly complex and therefore requires a more thorough examination [Wang et al., 2024].

To address these limitations, the concept of CoT prompting was introduced in Wei et al. [2023], demonstrating that instructing LLMs to reason step-by-step significantly improves performance on complex tasks. In this approach, LLMs are prompted to generate a series of short statements that mimic the logical process a person might use to solve a problem. Experiments revealed that CoT prompting enables models to achieve strong results in tasks of arithmetic, commonsense, and symbolic reasoning [Wei et al., 2023].

In an effort to enhance LLM reasoning abilities with CoT prompting, Wang et al. [2023b] introduces self-consistency to replace the single greedy decoding path in traditional CoT prompting [Wei et al., 2023]. Their method samples a variety of reasoning paths and identifies the most consistent answer by marginalizing across all possibilities [Wang et al., 2023b]. Beyond improving accuracy, this approach highlights the inherent diversity of reasoning paths within a single model, suggesting that multiple valid chains of reasoning can coexist.

Initiatives have also been put forward to extend and refine CoT prompting. Unlike traditional CoT where each step is independent, Least to Most prompting breaks difficult problems into sequential sub-problems where the outputs of previous steps are the inputs for the next [Zhou et al., 2023]. Moreover, recent efforts have examined the effects of partial or truncated CoT on model outputs. Lanham et al. [2023] measure faithfulness by truncating generated CoT at various points and re-prompting the model with only the partial reasoning.

Past research has indirectly measured the reasoning ability of LLMs by evaluating them on downstream reasoning tasks such as question answering or multi-hop inference [Huang and Chang, 2023]. Though, relying on the accuracy of the end task or the success rates is not indicative of step-by-step reasoning capability. Huang and Chang [2023] also explain that current performance measures mix reasoning ability with task knowledge, resulting in reasoning that cannot be evaluated in isolation. To resolve this, subsequent work [Nguyen et al., 2024] aims at reasoning process analysis directly, testing for logical coherence of individual steps, which provides more straightforward methods of reasoning quality evaluation.

LLMs frequently make errors when solving mathematical problems step-by-step, making it essential to identify where the errors occurred during the reasoning process [Zheng et al., 2025]. As a result, PRMs have been developed as a direct solution to the shortcomings of traditional indirect evaluation methods, which only assess final answers. PRMs are specifically designed to evaluate the correctness of each individual reasoning step, providing feedback that helps guide policy models toward more accurate and reliable mathematical reasoning [Zheng et al., 2025, Zhang et al., 2025]. PRMs output a score or probability that represents the model's confidence that the reasoning step is logically sound and contributes productively to problem resolution.

3 Methodology

We introduce a novel chain-splitting approach grounded in cumulative log-probability, whereby complete solutions are truncated at points of varying model confidence from an initial baseline model and then continued by a second continuation model. The methodology proceeds in three components: (Section 3.1) reasoning chain generation, (Section 3.2) chain truncation via cumulative log-probability, and (Section 3.3) model interchange protocols.

3.1 Reasoning Chain Generation

We use an initial model to generate complete reasoning chains for each problem in the test set. Each generation is performed with temperature set at 0.7, allowing a moderate degree of stochasticity in token sampling while still favoring high-probability continuations. Let the complete output chain be a sequence of tokens $r = \{t_1, t_2, \ldots, t_n\}$, with corresponding log-probabilities $\{\ell_1, \ell_2, \ldots, \ell_n\}$. We compute the cumulative log-probability up to position i as $L_i = \sum_{j=1}^{i} \ell_j$.

This sequence $\{L_1, L_2, \ldots, L_n\}$ defines the internal flow of confidence of the model throughout the reasoning process.

3.2 Chain Truncation via Log-Probability Thresholding

To identify semantically meaningful split points in the chain, we define three thresholds based on the total log-probability L_n:

- **25% truncation**: first index i such that $L_i \geq 0.25L_n$
- **50% truncation**: first index i such that $L_i \geq 0.50L_n$
- **75% truncation**: first index i such that $L_i \geq 0.75L_n$

For each threshold $\alpha \in \{0.25, 0.50, 0.75\}$, we extract the prefix $r_{1:k}$, where

$$k = \min\{i : L_i \geq \alpha L_n\}.$$

This results in three partially completed reasoning traces per problem, each grounded in the model's own internal confidence progression.

3.3 Model Interchange Protocol

Each truncated prefix is combined with a consistent CoT template meant for interchange that includes the original question, and the resulting prompt is provided to a secondary continuation model (further details in Section 8.1). We consider both intra-family and cross-family model pairings more precisely defined in Section 4.2. Each continuation model generates a single completion for each prefix using a temperature of 0.7, introducing controlled randomness to reflect typical sampling conditions while preserving coherence. These continuations are concatenated with the original prefix to form hybrid reasoning chains, which are then run through post-processing to extract the final answer using simple rule-based extraction.

All in all, for each problem instance, we obtain: one complete chain from the baseline generator, and multiple hybrid chains resulting from different continuation models and truncation depths (details in Section 4.2).

4 Experimental Setup

We now outline the experimental conditions under which our chain-splitting framework was evaluated. This includes: (Section 4.1) the dataset selected to benchmark reasoning difficulty and domain coverage, (Section 4.2) the models used for initial generation and continuation, and (Section 4.5) the metrics employed to quantify the quality of reasoning, compatibility, and the impact of performance on model interchanges.

4.1 Dataset Selection

An extensive dataset was carefully selected to capture a range of reasoning complexities and domain-specific scenarios.

- MATH [Hendrycks et al., 2021]: consists of 12,500 high-school and college-level mathematical problems that span diverse topics and demanding multi-step solutions, providing rigorous testing to evaluate advanced mathematical reasoning and generalization.

For our experiments, we evaluated models exclusively on the test splits of the MATH dataset, consisting of $5,000$ questions.

4.2 Model Selection and Configuration

We adopt Qwen2.5-PRM [Zheng et al., 2025] as our primary Process Reward Model, due to its fine-tuning on structured multi-step mathematical datasets such as PRM800K [Lightman et al., 2023] and Math-Shepherd [Wang et al., 2023a]. Qwen2.5-PRM is an instruction-tuned variant of Qwen2.5-Math-7B and supports token-level log-probability outputs.

For model interchange experiments, we select two baseline models and two continuation models:

4.3 Baseline

To establish a baseline for reasoning quality, we employ Gemma-3-4B-IT and LLaMA-3.1-70B-Instruct to generate CoT exemplars, two state-of-the-art instruction-tuned models from distinct architectural lineages.

- Gemma-3-4B-IT [Team et al., 2025], an instruction-tuned variant from the Gemma 3 model family developed by Google Deepmind with 4 Billion parameters is used to generate complete Chain-of-Thought reasoning paths, tuned to Gemma's architecture.
- LLaMA-3.1-70B-Instruct [Grattafiori et al., 2024], a large scale variant from the LLaMA 3 model family developed by Meta AI with 70 Billion parameters is used to generate complete Chain-of-Thought reasoning paths, tuned to LLaMA's architecture.

4.4 Continuation

- Gemma-3-1B-IT [Team et al., 2025], a lightweight variant from the same Gemma 3 model family, is used to evaluate how well reasoning chains can be completed by a structurally similar but smaller model.
- LLaMA 3.1-8B-Instruct Grattafiori et al. [2024] representing a different architectural lineage helps enable testing interchangeability across distinct LLM families. For brevity, we refer to the aforementioned models as Gemma and LLaMA respectively for the remainder of this paper.

On the MATH dataset, Gemma 3-1B-IT and Gemma 3-4B-IT performed with accuracies of 48.0% and 75.6% respectively [Team et al., 2025]. Moreover, Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct performed with accuracies 47.2% and 65.7% [Yang et al., 2024]. We observe that the two base models exhibit similar performance levels and, likewise, that the two continuation models perform comparably.

All models were prompted using one consistent CoT templates, either the interchange or full-run variant, as detailed in Section 8.1.

4.5 Evaluation Metrics

The hybrid reasoning chains generated were evaluated using a multifaceted set of metrics designed to assess accuracy, variability, and the impact of model interchanges on reasoning coherence and final outcomes. Specifically, we consider the following four core metrics:

- Answer Accuracy: Accuracy is defined as the proportion of final answers from generation that exactly match those from ground-truth solutions. This metric represents the model's ability to arrive at the correct final result through its reasoning chain.
- PRM Score: As a PRM is available for scoring, we additionally report average PRM-assigned scores that capture the internal likelihood and coherence of a given chain regardless of final correctness. We define the PRM score A' as the average plausibility score assigned to each reasoning step in a chain of n steps:

$$A' = \frac{1}{n}\sum_{i=1}^{n}\text{PRM}(s_i)$$

  where s_i denotes the i-th step in the chain. While traditional accuracy reflects outcome-level correctness, A' provides a step-level assessment of reasoning quality.
- Normalized Relative Gain (NRG): This metric quantifies whether incorporation of reasoning from another model helps or hinders performance. Given the accuracies of the original model A and B, and hybrid accuracies A' (Model A prefix + Model B suffix) and B' (Model B prefix + Model A suffix), we define:

$$\text{NRG}_A = \frac{A' - A}{A}, \quad \text{NRG}_B = \frac{B' - B}{B}.$$

  Positive values indicate a performance gain from model interchange, while negative values reflect degradation.
- Cross-Model Degradation (XMD): This metric captures the extent to which the continuation of a model degrades the original reasoning trajectory. It is defined as:

$$\text{XMD}_{A\to B} = \frac{A - B'}{A}, \quad \text{XMD}_{B\to A} = \frac{B - A'}{B}.$$

  XMD provides a normalized measure of reasoning incompatibility, where higher values indicate more severe disruptions introduced by the cross-model continuation.

5 Results

We present results across the MATH benchmark to evaluate model interchangeability across truncation points. Through our proposed metrics, we look to determine whether model continuation works to improve or disrupt the original reasoning trajectory. Experimental results were obtained using the Runpod cloud platform, leveraging NVIDIA H100 PCIe GPUs over approximately 250 GPU hours.

5.1 Full Chain-of-Thought Results

To establish a baseline, we first evaluate each model's performance using end-to-end CoT reasoning applied without interruption. For every example in the benchmark, the model is prompted to reason step by step to completion, producing a complete trajectory from question to final answer. We report results in terms of final answer accuracy and step-level reasoning score as seen in Table 1.

This baseline allows us to quantify native reasoning strengths and weaknesses of each model without the effects of interchange.

Model	Dataset	Accuracy (%)	PRM
Gemma-3-4B-IT	MATH	68.06%	0.8952
Gemma-3-1B-IT	MATH	36.28%	0.7904
LLaMA-3.1-70B-Instruct	MATH	60.80%	0.8725
LLaMA-3.1-8b-Instruct	MATH	47.76%	0.8522

Table 1: Performance of reasoning chains fully generated by each model (i.e., with no handoff or interchange from another model) on the MATH dataset.

5.2 Interchanged Chain-of-Thought Results

Thereafter, to gauge the interchangeability of reasoning processes across different models, we evaluate the completion of truncated CoT traces. Each reasoning chain is strategically truncated based on cumulative log-probability thresholds $(25\%, 50\%, 75\%)$, representing early, mid, and late points in the reasoning process. Subsequently, alternative models are assigned to continue the truncated reasoning chains through to completion.

We report performance for all continuation combinations, including accuracy, step-level scores, and coherence ratings as seen in Table 2 & Table 3. This analysis unveils the extent to which partial reasoning from one model can be reliably extended by another, highlighting cases of both successful handoff and systematic breakdowns that point to the limits of reasoning interchangeability.

Truncation	Continuation	Accuracy (%)	PRM	NRG	XMD
25%	Gemma-3-1B-IT	41.76%	0.7966	0.3678	0.3864
25%	LLaMA-3.1-8B-Instruct	43.60%	0.8393	0.3196	0.3594
50%	Gemma-3-1B-IT	49.86%	0.8002	0.3786	0.2674
50%	LLaMA-3.1-8B-Instruct	53.24%	0.8585	0.264	0.2177
75%	Gemma-3-1B-IT	55.26%	0.8032	0.3500	0.1881
75%	LLaMA-3.1-8B–Instruct	63.80%	0.8697	0.1853	0.0626

Table 2: Performance of hybrid reasoning chains by truncation point and continuation model on MATH dataset, using a fully generated CoT from Gemma-3-4B-IT.

Truncation	Continuation	Accuracy (%)	PRM	NRG	XMD
25%	Gemma-3-1B-IT	36.16%	0.7566	-0.1137	0.4053
25%	LLaMA-3.1-8B-Instruct	42.18%	0.8323	-0.0150	0.3062
50%	Gemma-3-1B-IT	38.50%	0.7730	-0.0968	0.3668
50%	LLaMA-3.1-8B-Instruct	46.26%	0.8456	-0.0072	0.2391
75%	Gemma-3-1B-IT	41.98%	0.7811	-0.0827	0.3095
75%	LLaMA-3.1-8B-Instruct	50.06%	0.8543	-0.0002	0.1766

Table 3: Performance of hybrid reasoning chains by truncation point and continuation model on MATH dataset, using a fully generated CoT from LLaMA-3.1-70B-Instruct.

6 Discussion

Our observations uncover degradation in performance when cross-family models are tasked to continue reasoning midstream initiated by a partially completed CoT. There are several factors likely responsible for this downgrade in performance:

6.1 Style and Representational Compatibility

A consistent disparity between intra-family and cross-family continuation highlights representational compatibility as a key factor in multi-model reasoning. Despite receiving high confidence chains, cross-family continuations (e.g., Gemma-3-4B-IT→LLaMA-3.1-8B-Instruct and LLaMA-3.1-70B-Instruct→Gemma-3-1B-IT) often fail to maintain correct reasoning. For instance, when LLaMA-3.1-70B-Instruct's chain is continued by Gemma-3-1B-IT, accuracy falls to 36.16% at the 25% mark-nearly a 40% relative decline compared to the base model's 60.80% full-chain accuracy-with a corresponding negative NRG (-0.1137). Similarly, continuations from Gemma-3-4B-IT into LLaMA-3.1-8B-Instruct under perform early on (43.60% at 25%) despite access to confident reasoning prefixes, yielding a lower NRG of 0.3196 compared to intra-family continuation at the same depth (Gemma-3-4B-IT→Gemma-3-1B-IT, 0.3678), indicating that these prefixes do not fully overcome differences in architecture and reasoning style. This pattern suggests a reasoning bias: each model family tends to rely more on its own reasoning patterns, which may result from structural differences between the families.

These results are consistent with previous work [Liu et al., 2023], which noted that structural differences between model families (GPT-4 in their case) can limit cross-model reasoning transfer, particularly for complex, multi-step reasoning tasks. While LLaMA models generate coherent chains within their own family, their internal reasoning representations differ from Gemma's, which may hinder smooth continuation across families. This is supported by consistently high XMD values across truncation points (e.g., 0.4053 at 25% and 0.3095 at 75% for LLaMA-3.1-70B-Instruct→Gemma-3-1B-IT), suggesting that reasoning coherence is not fully maintained even as longer prefixes are available. High-confidence reasoning prefixes do not appear sufficient to completely navigate these differences, indicating that cross-family continuation is constrained by family-specific reasoning tendencies.

In contrast, intra-family continuations show steady improvement with longer truncation depths. For example, when Gemma-3-1B-IT continues from Gemma-3-4B-IT, accuracy rises from 41.76% at 25% to 55.26% at 75%, accompanied by moderate NRG values ($0.3678 \rightarrow 0.3500$) and decreasing XMD ($0.3864 \rightarrow 0.1881$). Similarly when LLaMA-3.1-8B-Instruct continues from LLaMA-3.1-70B-Instruct, performance increases from 42.18% to 50.06%, with NRG improving from -0.0150 to near-neutral (-0.0002) and XMD decreasing from 0.3062 to 0.1766. These patterns suggest that the representational similarity between models supports a more stable continuation and better integration of context.

6.2 Context Integration Overhead

When deployed late in the reasoning chain (e.g., at the 75% mark), smaller continuation models such as Gemma-3-1B-IT and LLaMA-3.1-8B-Instruct must interpret and integrate extensive context generated by larger base models (Gemma-3-4B-IT and LLaMA-3.1-70B-Instruct). As reasoning sequences lengthen, models may face capacity limits that degrade performance. This bottleneck is attributed to the finite "working memory" of LLMs and the compounding demands of maintaining logical coherence across many steps [Shang et al., 2025]. The effect is especially pronounced when models are required to interpret and continue reasoning from an externally provided chain rather than generating all steps from scratch.

On the MATH dataset, truncation depth produces gradual improvements but does not eliminate the performance gap relative to non-handoff baselines. For example, when continuing Gemma-3-4B-IT's reasoning, Gemma-3-1B-IT improves from 41.76% at 25% truncation to 55.26% at 75%, while LLaMA-3.1-8B-Instruct rises from 43.60% to 63.80%. Similarly, when continuing LLaMA-3.1-70B-Instruct, LLaMA-3.1-8B-Instruct achieves a smoother progression from 42.18% to 50.06%, outperforming Gemma-3-1B-IT, which remains between 36.16% and 41.98%. These trends suggest that architectural alignment facilitates smoother context integration in same-family continuations, while representational mismatches in cross-family pairs disrupt coherent reasoning.

Despite improvements with longer prefixes, performance remains notably below that of fully self-generated chains (Gemma-3-4B-IT: 68.06%, LLaMA-3.1-70B-Instruct: 60.80%). XMD values confirm this persistent overhead: even at the 75% truncation point, XMD remains non-negligible (0.0626 for Gemma-3-4b-IT→LLaMA-3.1-8B-Instruct and 0.1766 for LLaMA-3.1-70B-Instruct→LLaMA-3.1-8B-Instruct), indicating incomplete recovery of original reasoning quality.

These observations highlight that truncation depth alone does not ensure effective reasoning transfer. Although larger prefixes reduce uncertainty and contextual loss, architectural and stylistic compatibility between base and continuation models remains the key factor determining success.

6.3 Error Amplification

Minor inconsistencies or ambiguities in early reasoning steps, especially when generated by a different model, can compound as LLaMA or Gemma continue the reasoning process. With limited steps remaining to revise earlier logic (particularly in final-answer-only completions), both models struggle to recover from upstream errors. These results suggest that effective interoperability in multi-step reasoning depends on both model capability and the degree of representational and contextual alignment across reasoning steps.

On the MATH dataset, when reasoning chains generated by Gemma-3-4B-IT or LLaMA-3.1-70B-Instruct are truncated and continued by smaller models at various points ($25\%, 50\%, 75\%$), performance declines in proportion to both truncation depth and cross-family divergence. When Gemma-3-4B-IT serves as the base, continuation by Gemma-3-1B-IT (intra-family) improves steadily from 41.76% at 25% to 55.26% at 75%, with NRG values rising from 0.3678 to 0.3500 and XMD decreasing from 0.3864 to 0.1881. Cross-family continuation by LLaMA-3.1-8B-Instruct performs competitively ($43.60\% \rightarrow 63.80\%$) but shows slightly lower NRG ($0.3196 \rightarrow 0.1853$), indicating weaker efficiency in utilizing the provided context. Longer prefixes appear to partially reduce representational mismatch, leading to more consistent performance over time.

When LLaMA-3.1-70B-Instruct serves as the base model, Gemma-3-1B-IT continuations perform substantially worse ($36.16\check{}41.98\%$ across truncation points) with persistently high XMD ($0.4053 \rightarrow 0.3095$) and negative NRG ($-0.1137 \rightarrow -0.0827$), suggesting limited transfer across families. Intra-family continuation by LLaMA-3.1-8B-Instruct performs more stably, reaching 50.06% at 75% with NRG improving from -0.0150 to -0.0002 and XMD decreasing from 0.3062 to 0.1766, reflecting more consistent reasoning integration within the same family.

Comparing fully generated chains with the hybrid results (Table 1), Gemma-3-4b and LLaMA-3.1-70B-Instruct still substantially outperform their continuations (68.06% and 60.80%, respectively). However, their smaller counterparts, especially Gemma-3-1B-IT, demonstrate partial to considerable recovery when inheriting sufficiently long prefixes, suggesting that similar architecture and tokenization structures enhance transfer performance. As seen over $25\%/50\%/75\%$ truncations, intra-family continuation (Gemma-3-4b→Gemma-3-1b) improves from $41.76\% \rightarrow 55.26\%$ ($+13.5$ pp), and even cross-family continuation (Gemma-3-4B-IT→LLaMA-3.1-8B-Instruct) exhibits a greater net gain of $43.60\% \rightarrow 63.80\%$ ($+20.2$ pp). In contrast, continuations from LLaMA-3.1-70b displayed weaker recovery with LLaMA-3.1-8B-Instruct rising only $+7.9$ pp (42.18% 50.06%), and Gemma-3-1B-IT gains just $+5.8$ pp ($36.16\% \rightarrow 41.98\%$). As truncation length increases, reasoning becomes more coherent, but full recovery is still unattainable, lending credence to how small representational gaps can compound through multi-step reasoning chains.

7 Conclusion

In this work, we introduced a novel framework for evaluating midstream interchangeability in large language models, grounded in a chain-splitting paradigm based on cumulative log-probability. By systematically truncating the reasoning chains generated by our base models and appending completions from either intra-family or cross-family models, we directly measured the stability and coherence of hybrid reasoning trajectories. Our experiments on MATH demonstrate that model family alignment plays a decisive role in the success or failure of such hybrid chains. While intra-family continuations generally preserved reasoning quality on simpler tasks, cross-family continuations often struggled to maintain coherence with the partial chains, despite comparable model performance as referenced in Section 4.2. This suggests that models like Gemma and LLaMA may be better aligned to continue reasoning within their own family than across different architectures.

These findings challenge previous assumptions about model modularity in contemporary NLP. Despite architectural advances and increasing performance parity across model families, our results suggest that inter-model transfer in multi-step reasoning remains fragile, constrained by differences in stylistic alignment, latent variable encoding, and contextual integration. The observed breakdowns reveal a significant gap between individual task performance and interoperability in reasoning, which is an area that has received insufficient attention in LLM evaluation.

More broadly, our work highlights the need for new approaches that preserve consistent semantic reasoning across different language models. As research advances toward compositional and multi-agent LLM systems, reliable interchangeability will become essential, not solely for efficiency, but also for alignment, verification, and interpretability. Our methodology provides an initial framework for diagnosing and quantifying this interoperability gap in a systematic, data-driven manner.

Limitations

- Single Completion Runs: All experiments were conducted using deterministic continuations. While this reflects realistic deployment scenarios, it limits our understanding of variance under sampling. Future work should evaluate robustness using multiple stochastic rollouts.
- Task Domain Scope: Our evaluation is confined to math reasoning (MATH). It remains unclear whether interchangeability generalizes to commonsense, scientific, or multimodal reasoning tasks.
- Domain-Specific PRMs: We employed a math-specific Process Reward Model (PRM). Evaluating reasoning quality in other domains will require retraining or adapting PRMs tailored to those reasoning distributions.

Future Work

- Cross-Domain Generalization: Evaluate model interchangeability on tasks such as commonsense QA, multi-hop retrieval, scientific explanation, and instruction-following, where reasoning formats may be more variable or implicit.
- Adaptive Truncation Strategies: Rather than using static log-probability thresholds $(25/50/75\%)$, future work could explore dynamic segmentation based on reasoning content, semantic shifts, or model uncertainty.
- Collaborative Model Architectures: Deploy multi-agent or multi-model reasoning pipelines in production environments (e.g., tutoring systems, scientific assistants) to study tradeoffs in latency, memory, and correctness.

8 Appendix

8.1 Prompting

Standardized Prompt

Full-Run Prompt :

System message: "You are a helpful assistant that solves problems step by step. Please provide clear reasoning with numbered steps and conclude with your final answer."

User message: "Solve this problem step by step:
Question: ['question']"

Interchange Prompt :

System message: "You are a helpful assistant that solves problems step by step. Please provide clear reasoning with numbered steps and conclude with your final answer."

User message: "Solve this problem step by step:
Question: ['question'] ['truncated reasoning']"

This prompt standardization ensures comparability in reasoning styles across models; slight variations were applied where necessary to accommodate model-specific tokenization or formatting requirements without altering the intended instructions or task semantics.

References

Xavier Amatriain. Prompt design and engineering: Introduction and advanced methods, 2024. URL https://arxiv.org/abs/2401.14423.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Konstantin Hebenstreit, Raphael Praas, Laura P. Kiesewetter, and Matthias Samwald. A comparison of chain-of-thought reasoning strategies across datasets and models. *PeerJ Computer Science*, 10: e1999, 2024. doi: 10.7717/peerj-cs.1999. URL https://doi.org/10.7717/peerj-cs.1999.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey, 2023. URL https://arxiv.org/abs/2212.10403.

Feihu Jin, Yifan Liu, and Ying Tan. Zero-shot chain-of-thought reasoning guided by evolutionary algorithms in large language models, 2024. URL https://arxiv.org/abs/2402.05376.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

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, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, May 2023. URL https://arxiv.org/abs/2305.20050.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. Logicot: Logical chain-of-thought instruction-tuning. *arXiv preprint arXiv:2305.12147*, May 2023. doi: 10.48550/arXiv.2305.12147. URL https://arxiv.org/abs/2305.12147.

Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs, 2024. URL https://arxiv.org/abs/2402.11199.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

HaoYang Shang, Xuan Liu, Zi Liang, Jie Zhang, Haibo Hu, and Song Guo. United minds or isolated agents? exploring coordination of llms under cognitive load theory. *arXiv preprint* *arXiv:2506.06843*, June 2025. doi: 10.48550/arXiv.2506.06843. URL https://arxiv.org/abs/2506.06843.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Chaojie Wang, Yanchen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. Q: Improving multi-step reasoning for llms with deliberative planning. *arXiv preprint arXiv:2406.14283*, June 2024.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, December 2023a. doi: 10.48550/arXiv.2312.08935. URL https://arxiv.org/abs/2312.08935.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023b. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning, 2025. URL https://arxiv.org/abs/2501.07301.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models, 2022. URL https://arxiv.org/abs/2210.03493.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning, 2025. URL https://arxiv.org/abs/2412.06559.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. URL https://arxiv.org/abs/2205.10625.