SOLAR: Scalable Optimization of Large-scale Architecture for Reasoning

Chen Li[1,*] Yinyi Luo[1,*] Anudeep Bolimera[1] Marios Savvides[1]

[1]Carnegie Mellon University, Pittsburgh, PA, USA

{chenli4, yinyil, abolimer, marioss}@andrew.cmu.edu

[*]Equal contribution

Abstract

Large Language Models (LLMs) excel in reasoning but remain constrained by their Chain-of-Thought (CoT) approach, which struggles with complex tasks requiring more nuanced topological reasoning. We introduce **SOLAR** (Scalable Optimization of Large-scale Architecture for Reasoning), a framework that dynamically optimizes reasoning topology—including trees and graphs—to enhance accuracy and efficiency.

Our Topological-Annotation-Generation (TAG) system automates topological dataset creation and segmentation, improving post-training and evaluation. Additionally, we propose Topological-Scaling, a reward-driven framework that aligns training and inference scaling, equipping LLMs with adaptive, task-aware reasoning.

SOLAR achieves substantial gains on MATH and GSM8K: **+5%** accuracy with Topological Tuning, **+9%** with Topological Reward, and **+10.02%** with Hybrid Scaling. It also reduces response length by over **5%** for complex problems, lowering inference latency.

To foster the reward system, we train a multi-task Topological Reward Model (M-TRM), which autonomously selects the best reasoning topology and answer in a single pass, eliminating the need for training and inference on multiple single-task TRMs (S-TRMs), thus reducing both training cost and inference latency. In addition, in terms of performance, M-TRM surpasses all S-TRMs, improving accuracy by **+10%** and rank correlation by **+9%**.

To the best of our knowledge, SOLAR sets a new benchmark for scalable, high-precision LLM reasoning while introducing an automated annotation process and a dynamic reasoning topology competition mechanism.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable advancements in natural language processing, particularly in complex reasoning tasks. Despite their success, LLMs primarily rely on a sequential chain-of-thought (CoT) reasoning structure by default. Regardless, many real-world problems require more intricate topological reasoning structures, such as trees or graphs, to achieve optimal solutions. In this work, we propose **SOLAR** (Scalable Optimization of Large-scale Architecture for Reasoning), a framework that dynamically optimizes reasoning pathways to enhance an LLMs performance.

1.1 Observations on LLMs Reasoning Patterns

Through systematic evaluation, we observed the following key phenomena in LLM-based reasoning:

- LLMs predominantly default to CoT reasoning and seldom generate more sophisticated reasoning topologies, such as tree-of-thought or graph-of-thought, without explicit prompting.
- Certain complex problems, including but not limited to two-group matching problems, the Travelling Salesman Problem (TSP), and multi-stage robotic manipulation struggle with the default CoT reasoning and exhibit performance gains when alternative topological reasoning structures are employed.

1.2 Our Approach

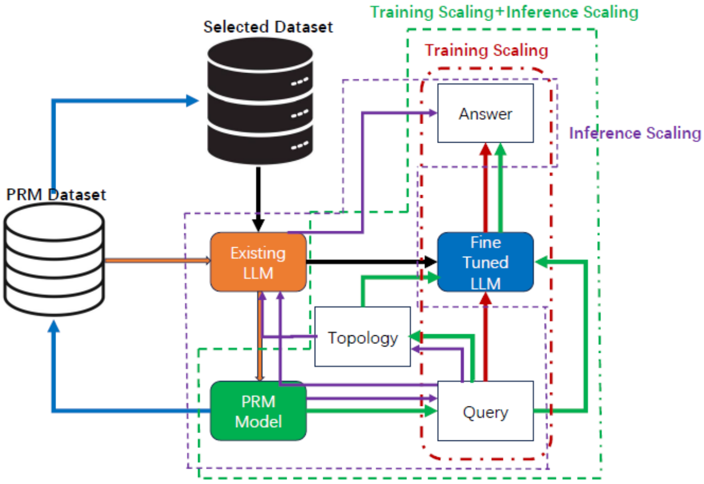

Figure 1: Overall Framework

To address these observations, we hypothesize that different reasoning problems have distinct optimal reasoning topologies that can yield higher accuracies. To validate this hypothesis and enhance an LLMs' reasoning performance, we developed a multi-stage pipeline:

Evaluation Pipeline We systematically analyzed three reasoning topologies—Chain-of-Thought (CoT), Tree-of-Thought (ToT), and Graph-of-Thought (GoT)—using models of varying capacities on two mathematical reasoning datasets, MATH and GSM8K. Our evaluation led to three key insights:

- Different problems favor different topological structures, with each topology exhibiting distinct advantages in maximizing accuracy.
- The overall accuracy of ToT and GoT remains comparable to CoT, suggesting that less frequently generated reasoning topologies do not suffer from performance degradation.
- These patterns persist across both smaller LLMs and state-of-the-art (SOTA) reasoning models, highlighting their robustness and generalizability.

Synthetic Topological Data Infrastructure To enable scalable research, we developed an automated system for generating and annotating reasoning datasets with diverse topological structures. This infrastructure enables a more objective segmentation of problems by difficulty level, mitigating human biases inherent in heuristic-based labeling and difficulty classification—a common challenge in constructing controllable synthetic datasets.

Topological-Scaling Framework We introduce the Topological-Scaling Framework, which operates as a competitive selection process where the optimal reasoning structure and the best answer for each problem are dynamically determined at test time (i.e., a winner identification mechanism). To achieve this, we design a hierarchical pipeline that integrates post-training with inference-time rewarding and optimization, enabling adaptive topology selection.

- **Topological Tuning**: A training-scaling approach that fine-tunes a base LLM to learn how to generate optimal reasoning topology policy, yielding a **+5%** accuracy improvement while maintaining relatively low inference latency—making it an efficient choice for real-time applications. Notably, our experiments reveal that for more complex tasks, such as **MATH**, topological tuning reduces the token length of generated responses by **5%**, further lowering inference latency.

- **Topological Rewarding**: An inference-scaling method leveraging an in-house trained multi-task topological reward model (M-TRM) to dynamically select the best reasoning topology and the winning answer at test time. This strategy shifts computational cost from training to inference, prioritizing accuracy gains, enhancing accuracy, with **+9%** accuracy boost, at the expense of higher inference latency. M-TRM autonomously selects the best reasoning topology and answer in a single pass, eliminating the need for multiple single-task models. M-TRM surpasses all single-task TRMs, improving accuracy by **+10%** and rank correlation by **+9%**.
- **Hybrid Scaling**: A unified approach that integrates both training scaling and inference scaling to maximize performance gains. This method not only learns to generate optimal policies but also further refines candidate responses that have already undergone a selection process based on competitive filtering. It achieves the highest accuracy improvement of **+10.02%**, albeit with increased computational demands and higher inference latency.

We conduct comprehensive experiments to evaluate the impact of all three approaches, demonstrating their respective trade-offs in efficiency, computational cost, and generation quality for downstream tasks. An overview of our architecture is shown in Figure 1.

.

1.3 Contributions

Our work makes the following key contributions:

- **Empirical Insights into Topological Reasoning**: We conduct a systematic evaluation demonstrating that different reasoning problems benefit from distinct reasoning topologies, validated across multiple datasets and model sizes.
- **Topological-Annotation-Generation System (TAG)**: We develop an advanced data infrastructure that automates the generation and annotation of synthetic reasoning datasets with diverse topologies. TAG serves as a crucial asset for post-training and facilitates scalable research in this domain. Additionally, it enables objective task difficulty categorization and answer quality assessment, reducing heuristic human biases and allowing for a more fine-grained analysis of problem subsets.
- **Hierarchical Topological-Scaling Framework**: We introduce a novel Topology Competition Mechanism that unifies training-scaling and inference-scaling optimizations to significantly enhance LLM reasoning performance. This framework not only delivers substantial performance gains but also provides adaptive selection

strategies, enabling flexible trade-offs between accuracy, computational cost, and efficiency.

Our results demonstrate unprecedented improvements on **MATH** and **GSM8K**, highlighting the effectiveness of learning-based adaptive rewarding in topological reasoning for enhancing LLM performance while ensuring efficiency.

2 Related Work

2.1 Reward Models in LLMs Reasoning

Reward Models (RMs) have been widely adopted in the training and inference of Large Language Models (LLMs) to guide reasoning and improve response quality [Christiano et al., 2023]. Two primary types of RMs have been explored: Outcome Reward Model (ORM) and Process Reward Model (PRM). While ORM assigns rewards based on the correctness of final outputs, such as those used in RLHF[Ouyang et al., 2022] and RLAIF [Bai et al., 2022], PRM evaluate intermediate reasoning steps, allowing for a more structured and interpretable learning process.

In recent years, PRM have gained attention in reasoning models due to their ability to enhance generation quality through inference scaling. By reinforcing reasoning pathways step-by-step, PRM enable LLMs to dynamically adjust to complex problem-solving tasks [Lightman et al., 2023]. B-STaR [Zeng et al., 2024] optimizes self-taught reasoners by balancing exploration and exploitation in iterative learning using external rewards. These advancements have led to improved reasoning efficiency, particularly in multi-step and structured problem domains.

2.2 Scaling Laws in Large Language Models

Scaling laws [Shuai et al., 2024] are fundamental to LLMs evolution, impacting both training and inference performance. Training scaling improves accuracy by increasing model size and data but is computationally costly. Inference scaling, in contrast, optimizes reasoning at test time by integrating ”slow thinking” and ”fast thinking” [Lightman et al., 2023], balancing computational depth and efficiency. Studies [Wu et al., 2024] highlight its role in structured problem-solving. The work of [Ma et al., 2023] enhances the reasoning of LLMs using step-level rewards and heuristic search to improve multi-step problem solving.

2.3 Advances in Topological Reasoning

Traditional LLMs rely on Chain-of-Thought (CoT) reasoning [Wei et al., 2023], following a linear sequence. Recent advances introduce Tree-of-Thought (ToT) [Yao et al., 2023] and Graph-of-Thought (GoT) [Besta et al., 2024] for more complex decision-making. ToT enables hierarchical reasoning, while GoT allows interconnected paths, excelling in tasks like TSP and robotics. Despite their benefits, existing methods heuristically predefine reasoning topologies when solving problems, overlooking that optimal structures vary by problem and must be dynamically learned for optimal performance.

2.4 Curriculum Learning for Structured Reasoning

An emerging area relevant to topological reasoning is *curriculum learning*, where training data are structured and progressively introduced to align with the model's evolving capabilities [Bengio et al., 2009]. By gradually increasing task complexity, curriculum learning reinforces structured problem-solving skills and enhances LLM reasoning performance. Recent studies demonstrate its effectiveness in various domains: Xi et al. [Xi et al., 2024] propose a *reverse curriculum reinforcement learning* approach, where LLMs are trained on progressively easier problems before tackling complex ones, leading to improved reasoning efficiency. Zhao et al. [Zhao et al., 2024] introduce *Automatic Curriculum Expert Iteration (Auto-CEI)*, which refines reasoning through iterative curriculum-based self-training, improving reliability in complex tasks. Ma et al. [Ma et al., 2025] develop a *Problem-Solving Logic Guided Curriculum*, which leverages problem-solving heuristics to enhance LLM in-context learning for complex reasoning. This stream of work is orthogonal to ours, and when combined with rewarding-based topological reasoning, curriculum learning provides a powerful mechanism for optimizing both training efficiency and inference performance in scalable LLMs. We leave this area for future research.

To the best of our knowledge, we are the first to systematically explore the impact of diverse topological structures on LLMs' reasoning and to design a unified framework that integrates post-training paradigms with inference scaling, leveraging the unique strengths of adaptive learning-based topological reasoning. This breakthrough redefines scalable LLMs' optimization, unlocking new frontiers in complex problem-solving.

3 Methodology

3.1 Hypothesis Validation and Evaluation Methods

3.1.1 Observations and Hypothesis

We begin by analyzing the reasoning patterns of LLMs when solving mathematical problems. Through systematic evaluation, we observe the following phenomena:

- LLMs predominantly generate *chain-of-thought* (CoT) reasoning and seldom adopt more complex reasoning structures such as *tree-of-thought* (ToT) or *graph-of-thought* (GoT).
- Certain problems, such as *Data Center Fault Tolerance*, the *Traveling Salesman Problem (TSP)*, and *Multi-stage Robotic Manipulation*, benefit from richer topological reasoning structures beyond CoT.

Based on these observations, we propose the following two hypotheses to be validated in later sections:

- **Hypothesis 1:** Different problems require distinct optimal reasoning topologies that yield the best solutions.
- **Hypothesis 2:** Solving problems with optimal topological reasoning structures can significantly enhance generation accuracy.

3.1.2 Validating Hypothesis 1: Topological Annotation and Evaluation

To validate **Hypothesis 1**, we designed and implemented an automated data generation and annotation system, the Topological-Annotation-Generation (TAG) System (detailed in Section 3.2.1). This system constructs a synthetic dataset where each sample consists of: (1) a problem statement paired with a group of generated responses, (2) multiple reasoning topologies, including CoT, ToT, and GoT, and (3) a hierarchical labeling system annotated automatically.

Specifically, this hierarchical labeling system is illustrated as below. Each sample in the dataset is automatically annotated with two labels:

- **Topo Label:** A continuous value in the range $[0, 1]$, representing the probability that a given topology produces the correct answer for a question.

- **Hard Label:** A binary value $\{0, 1\}$, indicating whether the generated answer is correct.

With these labels, we evaluate each reasoning topology by defining the following two metrics:

- **Accuracy:** The proportion of correct answers generated using each topology.
- **Win Rate:** The likelihood of each topology being the best-performing structure across all questions.

Win Rate Calculation The **Win Rate** of a topology $T \in \{CoT, ToT, GoT\}$ is defined as:

$$\text{WinRate}(T) = \frac{|\{q \in Q \mid T = \arg\max_{T' \in \{CoT, ToT, GoT\}} \text{Topo-label}(q, T')\}|}{|Q|} \tag{1}$$

where Q is the total set of questions, and Topo-label(q, T) denotes the topo-label of topology T for question q. For each question, the topology with the highest topo-label is assigned a win. The win rate for each topology is then computed as the fraction of questions where it was optimal.

Experimental results (detailed in Section 4.2) confirm that different problems exhibit different optimal topological reasoning structures, a phenomenon agnostic to model size or capacity, thus validating Hypothesis 1.

3.1.3 Validating Hypothesis 2: Performance Boost With Topological Scaling

To validate **Hypothesis 2**, we design and implement a hierarchical, adaptive-learning-based rewarding framework, Topological Scaling, which harnesses the synergy between training-scaling and inference-scaling in a multi-topological reasoning space. We conduct rigorous ablation studies to evaluate the impact of our approach.

Experimental results (presented in Section 4) demonstrate significant performance improvements, further supporting the Hypothesis 2. The details of our methodology are illustrated in Section 3.3.

3.2 Synthetic Topological Data Infrastructure

3.2.1 Topological-Annotation-Generation System (TAG)

In this section, we outline our approach in automatically annotating the topology reasoning dataset. We begin by introducing the datasets used in our study, followed by a detailed breakdown of data generation and annotation process.

Datasets This work leverages two datasets: GSM8K [Cobbe et al., 2021] and MATH [Hendrycks et al., 2021]. For training purpose, we split both datasets to training and testing sets. The final constructed synthetic data can be used for both post-training purpose and for evaluation purpose.

Data Generation To ensure diversity in reasoning topologies and a balanced distribution of positive and negative samples in our dataset, we utilized both a small-scale model, Qwen2-VL-7B-Instruct [Wang et al., 2024], and an open-source state-of-the-art reasoning model with hundreds of billions of parameters. These models generated responses across three reasoning topologies—Chain-of-Thought (CoT), Tree-of-Thought (ToT), and Graph-of-Thought (GoT)—with extensive degree of freedom in maximum depth, number of children, and number of neighbors.

Automatic Annotation As described in Section 3.1.2, we assign each problem a Topo Label and each response a Hard Label.We design an automated annotation pipeline for topological reasoning as follows:

First, using the generation mechanism outlined in the paragraph above, we obtain a diverse set of responses for each question, covering all three reasoning topologies—Chain-of-Thought, Tree-of-Thought, and Graph-of-Thought. We then apply the following annotation process:

- Topo Label ($\mathcal{T}_q$): This is a task-specific label which represents how effective each reasoning topology solves a given problem. For each problem q, we first compute the accuracy of responses generated by each reasoning topology, and then assign the accuracy to each reasoning topology as its Topo Label:

$$\mathcal{T}_q = \max_{T \in \{\text{CoT, ToT, GoT}\}} \frac{N_{\text{correct}}(q, T)}{N_{\text{total}}(q, T)} \tag{2}$$

where $N_{\text{correct}}(q, T)$ is the number of correct responses using topology T for question q, and $N_{\text{total}}(q, T)$ is the total number of responses generated using T. The resulting $\mathcal{T}_q$ is a continuous value in $[0, 1]$.

- Hard Label ($\mathcal{H}_a$): This is a response-specific label which is a variant of a binary Outcome-Reward-Model(ORM) label. Each response a is assigned a 1 if correct and 0 if incorrect:

$$\mathcal{H}_a = \begin{cases} 1, & \text{if } a \text{ is correct} \\ 0, & \text{if } a \text{ is incorrect} \end{cases} \tag{3}$$

These annotations allow us to quantitatively evaluate the performance of different reasoning topologies and assess their impact on problem-solving accuracy.

3.2.2 Problems Difficulty Segmentation

With **TAG**, we gain an additional advantage: the ability to analyze problems from an entirely new perspective. By examining the distribution of topological (topo) labels across all three reasoning structures, we can redefine problem difficulty in a data-driven manner, reducing reliance on heuristic biases. Specifically, we categorize problems as follows:

- **Hard**: Problems where all three topo labels fall below a specified quantile threshold in their respective distributions.
- **Easy**: Problems where all three topo labels exceed a specified quantile threshold in their respective distributions.
- **Medium**: Problems that do not fall into either the hard or easy categories.

This classification system provides a toolkit for further finer-grained research.

3.3 Topological Scaling for Enhanced Reasoning

Topological Tuning We perform Supervised Fine-Tuning (SFT) using high-quality topological reasoning data selected by TAG, which is split into train and test sets. Training data is selected through the following three-step process:

- **Diversity Sampling**: To ensure a balanced dataset, we sample the same proportion of data from hard, easy, and medium problems, respectively, based on the difficulty segmentation definition in Section 3.2.2.

- **Correct Answer Filtering**: We only retain only correct responses, which have hard label.
- **Rejection Sampling (RS)**: Following [Grattafiori et al., 2024, Qwen et al., 2025], we apply RS with an in-housed trained topological reward model to filter out spurious samples. The reward model is detailed in the next paragraph.

We then train the model using Next Token Prediction on this curated dataset. The base model for Supervised Fine-Tuning (SFT) is Qwen2-VL-7B-Instruct [Wang et al., 2024], with fine-tuning performed using LoRA [Hu et al., 2021] for parameter-efficient adaptation.

This training-scaling strategy is optimized for real-time applications that demand low inference latency and high accuracy. As shown in Section 4.3, fine-tuning the model with diverse topological reasoning data surpasses the baseline, producing shorter yet more accurate responses, ultimately reducing latency.

Topological Rewarding At inference time, we introduce a Topological Rewarding mechanism to select the best response among multiple reasoning topologies—a process we refer to as the Topology Competition Game. Given a problem, a base model (with or without fine-tuning) first generates a diverse set of responses using different reasoning topologies. We then employ our in-house trained multi-task Topological Reward Model (M-TRM) to identify the optimal reasoning topology at the problem level and the best response at the answer level in a single pass. We provide two usage scenarios for this system: 1) Inference-Scaling Only: If the response-generation policy model is not fine-tuned, the approach relies solely on inference scaling. 2) Hybrid Scaling: If the policy model is fine-tuned (e.g., via topological tuning), the approach integrates both training scaling and inference scaling effects, as detailed in the next paragraph. Experimental results are presented in Section 4.4.

Hybrid Scaling This follows Scenario 2 described above, where the generation policy model is a topologically tuned model. This approach seamlessly integrates training-scaling with inference-scaling, achieving the highest performance gains. However, it requires increased computation during both training and inference, leading to higher latency. This strategy is best suited for downstream tasks that align with its performance objectives and computational constraints. Experimental results are presented in Section 4.4.

4 Experiments

4.1 Experiment Setup

In this section, we evaluate the effectiveness of our proposed method on challenging mathematical problems that require complex reasoning. We use two benchmark datasets, GSM8K and MATH, as our base datasets and apply TAG to generate a synthetic topological reasoning dataset annotated with both Topo Labels and Hard Labels. To facilitate downstream post-training, including SFT and M-TRM, we split the dataset into training and test sets. All experiments are conducted on eight NVIDIA A100 GPUs.

For evaluation, we use Accuracy and Win Rate as defined in Section 3.1.2. We assess these metrics topology-wise and also compute an Overall Accuracy, which reflects performance across samples from all reasoning topologies.

The structure of this section is as follows: Section 4.2 presents our evaluation results for validating Hypothesis 1, as introduced in Section 3.1.2. Sections 4.3 and 4.4 validate Hypothesis 2, also proposed in Section 3.1.2. Specifically, Section 4.3 examines the impact of Topological Tuning, including an ablation study, while Section 4.4 evaluates the effects of Topological Rewarding and Hybrid Scaling. Finally, Section 4.5 discusses the trade-offs between performance and efficiency across the three scaling strategies provided by our framework with flexibility.

4.2 Topological Reasoning Validation

In this section, we validate Hypothesis 1 proposed in Section 3.1.2

To test our hypothesis, we select Qwen2-VL-7B-Instruct as our base model [Wang et al., 2024] due to its superior ability to reason beyond linear chain structures. We evaluate the success rates of generating tree and graph reasoning structures across different models. To ensure robustness, each model is prompted to generate a specific reasoning topology for 100 questions, with five trials per topology.

Results in Figure 2 show that Qwen2.5-Math [Yang et al., 2024] achieves success rates of 11%, while another leading mathematical reasoning model achieves 7%. In contrast, Qwen2-VL-7B-Instruct [Wang et al., 2024] significantly outperforms both, achieving a 74% success rate, making it the ideal base model for our study.

We hypothesize that Qwen2-VL's superior multi-topology reasoning generation stems from its exposure to high-dimensional spatial information during pre-training, potentially

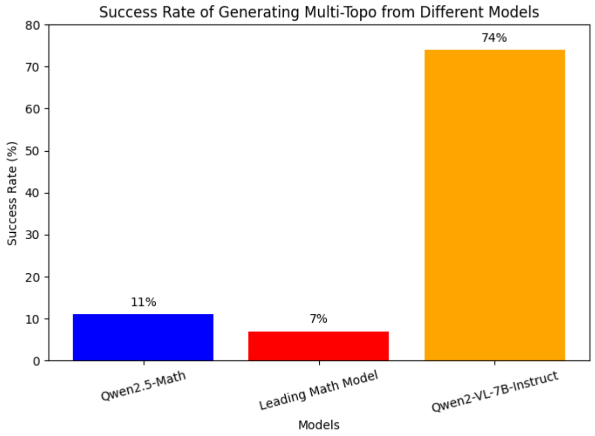

Figure 2: Success Rate of Genrating Multi-topology from different

enhancing its ability to encode non-Euclidean structures. This unique capability warrants further investigation.

As discussed in Section 3.1.2, we present the Win Rate and Accuracy for each reasoning topology to validate Hypothesis 1. Figure 3 demonstrates that less commonly generated structures—such as tree and graph—perform on par with the chain topology in terms of accuracy. However, notably, Figure 4 shows different problems favor different reasoning topologies, underscoring the potential advantages of incorporating trees and graphs into the reasoning process. Furthermore, we observe that these trends persist across models of varying scales, from 7B parameters to state-of-the-art reasoning models with hundreds of billions of parameters, indicating their universality and agnosticism to model size and capacity. Thus, our findings confirm Hypothesis 1. More detailed numbers are in Table Table 2 in Appendix.

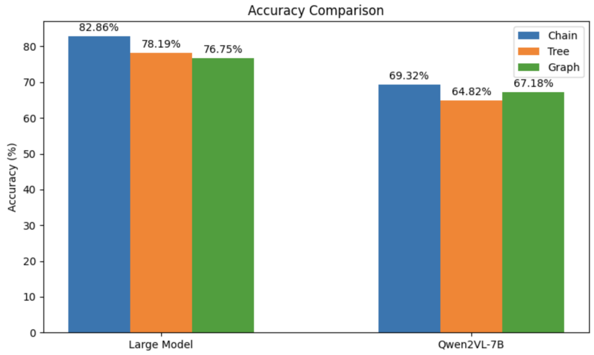

Figure 3: Accuracy comparision with existing models

In the next two sections, Section 4.3 and Section 4.4, we will validate Hypothesis 2 proposed in Section 3.1.2

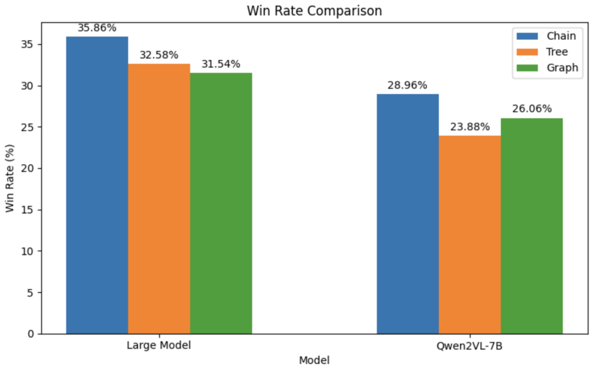

Figure 4: Win rate with existing models

4.3 Topological Tuning Impact

4.3.1 Topological Tuning Results

We fine-tuned Qwen2VL-7B-Instruct model using TAG annotated data following a selection process illustrated in Section 3.3 data, which is mixed with alpaca dataset [for Research on Foundation Models , CRFM] to prevent catastrophic forgetting.

To evaluate the performance, we test our fine-tuned model on the out-of-sample test set. Topological Tuning results are presented in Figure 5. The **+5%** improvement in accuracy underscores the impact of training with structured, high-quality data, while also demonstrating the effectiveness of the TAG mechanism in generating, annotating, and selecting relevant examples. This enhances complex reasoning capabilities and improves problem-solving accuracy. Additionally, the **5%** reduction in generated token length highlights the potential for achieving higher accuracy with lower inference latency.

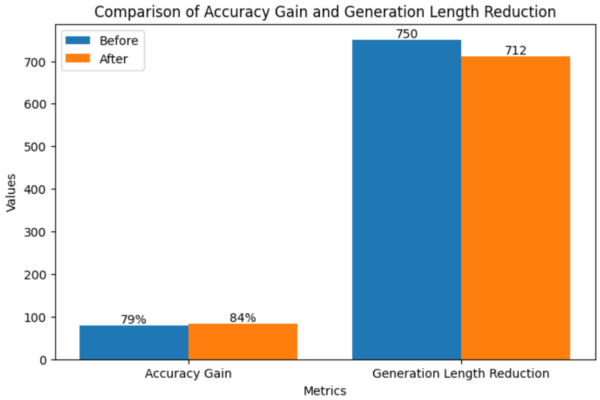

Figure 5: Topological Tuning Results Overall

We further compare our fine-tuned model with the baseline by explicitly prompting it to reason with all three reasoning topologies. The topology-wise Accuracy and Win Rate before and after Topological Tuning are shown in Figure 6 7 Detailed numbers are show in the Table 3 in Appendix.

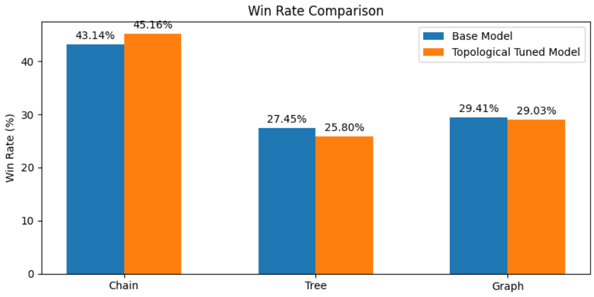

Figure 6: Topo-wise Win Rate Comparison

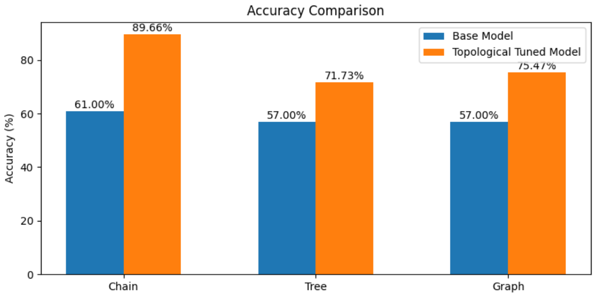

Figure 7: Topo-wise Accuracy Comparison

4.3.2 Ablation Study

To rule out the possibility that performance gains are solely due to the fine-tuning process rather than the multi-topology effect, we conduct an ablation study to assess the additional value provided by diverse topological reasoning. We compare a model trained exclusively on chain topology data—the default behavior of most state-of-the-art reasoning models—against a model trained on a mix of all three reasoning topologies. Both models are trained on the same number of samples.

Results, shown in Figure 8, indicate that the multi-topology model outperforms the chain-only model in overall accuracy, CoT accuracy, and GoT accuracy, while exhibiting a slight drop in ToT accuracy. Since these findings confirm that learning from optimal reasoning topologies improves overall accuracy—and given that variations across individual reasoning topologies are expected—the minor decline in ToT performance is acceptable and does not invalidate our main hypothesis. More detailed numbers are in the Table 4 in Appendix.

1. Ablation Observation: Overall Accuracy

- Fine-tuning solely on chain reasoning improves accuracy, even with high-quality filtered data (Section 3.3).

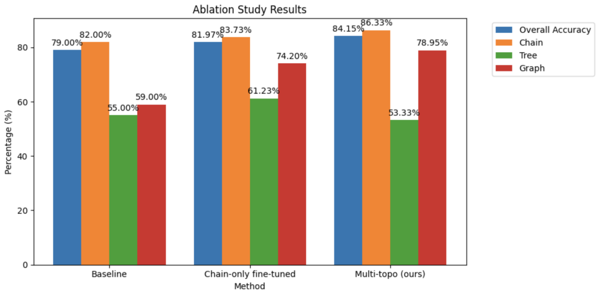

Figure 8: Ablation study

- Incorporating tree and graph reasoning further enhances performance, demonstrating the added value of **diverse topological tuning**, eliminating the possibility that the gain reported in Section 4.3.1 is just due to training from high-quality data. Instead, the results demonstrate robust gain from diverse topological reasoning injection.

2. Bonus Insights: Accuracy Across Different Topologies Mixing topologies creates a **synergistic effect**: training on chain reasoning improves graph reasoning accuracy, while a mixed approach enhances chain reasoning. These findings suggest mutual benefits across topologies, highlighting an open research direction to further optimize LLMs' reasoning.

Results shown in Figure 6, Figure 7, and Figure 8 collectively demonstrate the effectiveness of Topological Tuning.

4.4 Impact of Topological Rewarding and Hybrid Scaling

In this section, we evaluate the effectiveness of Topological Rewarding, an inference-scaling-only strategy, and Hybrid Scaling, which combines both training scaling and inference scaling.

For Topological Rewarding, we use the Qwen2-VL-7B-Instruct model without fine-tuning as the generation policy model. We then apply our in-house trained multi-task Topological Reward Model (M-TRM) to determine the optimal reasoning topology and select the best response for each problem at test time. Finally, we compare accuracy with and without Topological Rewarding to assess its impact.

For Hybrid Scaling, we apply Topological Rewarding on top of a topologically tuned model, following the same process as described above.

Results are presented in Figure 9. At a high level, the training-scaling strategy (Topological Tuning) achieves a **+5%** accuracy improvement, the inference-scaling-only strategy

(Topological Rewarding) yields a **+9%** gain, and Hybrid Scaling, which seamlessly integrates both approaches, achieves the highest improvement at **+10.02%**.

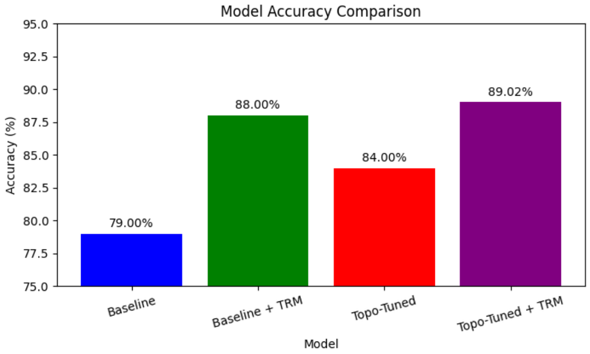

Figure 9: Hybrid Scaling

4.5 Discussion

Table 1: Accuracy Results for Three Scaling Comparison

Method	Overall Accuracy	Test Latency
Baseline	79%	Medium
Topo-Rewarding (ours)	88%	High
Topo-Tuning (ours)	84%	Low
Hybrid-Scaling (ours)	89.02%	Medium to High

From the results in Table 1, we observe that all three approaches—Topological Rewarding, Topological Tuning, and Hybrid Scaling—enhance performance to varying degrees while exhibiting different inference latencies. Topological Tuning and Hybrid Scaling are particularly beneficial in low- to medium-latency scenarios, as the Topological Tuning mechanism reduces generation length for more complex tasks, thereby lowering inference time.

We hypothesize that this phenomenon arises because the model, by adaptively learning from a diverse set of "winning" and "losing" reasoning topologies across problems of varying complexity, develops a greater resilience to overthinking. A more rigorous study into the underlying rationale behind this effect is left for future research.

This flexibility allows downstream tasks to select the most suitable method based on specific objectives and constraints, such as accuracy, inference latency, and computational resources.

5 Conclusion

We introduce SOLAR, a paradigm shift in LLM reasoning that adaptively learns and dynamically identifies the optimal reasoning topology—chain, tree, or graph—along with the best response for each task. SOLAR employs a unified approach that integrates both training scaling and inference scaling to maximize performance gains. This method not only learns to generate optimal policies but also refines candidate responses that have already undergone a competitive selection process.

Powered by Topological-Annotation-Generation (TAG) and Topological Scaling, SOLAR significantly outperforms the baseline on both MATH and GSM8K datasets. A comprehensive experimental analysis, supported by an ablation study, further demonstrates the robustness of our approach, validating our hypothesis on adaptive reasoning. This enables SOLAR-powered LLMs to break free from the constraints of the traditional default Chain-of-Thought (CoT) reasoning process.

Furthermore, the observed reduction in response length following Topological Tuning for complex tasks suggests that our framework offers greater flexibility for downstream tasks, enabling a more effective trade-off between performance, efficiency, and computational cost. This phenomenon, which we refer to as ”resilience to overthinking,” merits further investigation and is left for future research.

Our findings further highlight the **synergistic effect** between diverse topologies, revealing that reasoning pathways are not a one-size-fits-all solution but should be task specific and inherently tailored according to the complexity of the problem. The observed performance divergence across different problem segments further points towards hidden structural properties that influence an LLMs reasoning efficiency.

Our work paves the way for various research directions where we propose the following open ended questions: Which underlying mechanisms drive the performance divergence in different problem segments? How can the observed **synergistic effect** between reasoning structures and scaling laws be further optimized? Can dynamic topological reasoning generalize to broader domains beyond mathematical problem-solving? Addressing these questions will not only deepen our understanding of LLM cognition but also unlock new frontiers in adaptive reasoning architectures, paving the way for more scalable, efficient, and ethical AI systems.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL `https://arxiv.org/abs/2204.05862`.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL `https://dl.acm.org/doi/10.1145/1553374.1553380`.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16): 17682–17690, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i16.29720. URL `http://dx.doi.org/10.1609/aaai.v38i16.29720`.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL `https://arxiv.org/abs/1706.03741`.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL `https://arxiv.org/abs/2110.14168`.

Stanford Center for Research on Foundation Models (CRFM). Alpaca: A strong, open-source instruction-following model, 2023. URL `https://crfm.stanford.edu/2023/03/13/alpaca.html`. Accessed: YYYY-MM-DD.

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.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

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, 2023. URL https://arxiv.org/abs/2305.20050.

Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. Let's reward step by step: Step-level reward model as the navigators for reasoning, 2023. URL https://arxiv.org/abs/2310.10080.

Xuetao Ma, Wenbin Jiang, and Hua Huang. Problem-solving logic guided curriculum in-context learning for llms complex reasoning, 2025. URL https://arxiv.org/abs/2502.15401.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Xian Shuai, Yiding Wang, Yimeng Wu, Xin Jiang, and Xiaozhe Ren. Scaling law for language models training considering batch size, 2024. URL https://arxiv.org/abs/2412.01505.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

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.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models, 2024. URL https://arxiv.org/abs/2408.00724.

Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, Honglin Guo, Wei Shen, Xiaoran Fan, Yuhao Zhou, Shihan Dou, Xiao Wang, Xinbo Zhang, Peng Sun, Tao Gui, Qi Zhang, and Xuanjing Huang. Training large language models for reasoning through reverse curriculum reinforcement learning, 2024. URL https://arxiv.org/abs/2402.05808.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL `https://arxiv.org/abs/2305.10601`.

Weihao Zeng, Yuzhen Huang, Lulu Zhao, Yijun Wang, Zifei Shan, and Junxian He. B-star: Monitoring and balancing exploration and exploitation in self-taught reasoners, 2024. URL `https://arxiv.org/abs/2412.17256`.

Zirui Zhao, Hanze Dong, Amrita Saha, Caiming Xiong, and Doyen Sahoo. Automatic curriculum expert iteration for reliable llm reasoning, 2024. URL `https://arxiv.org/abs/2410.07627`.

Appendix

A Additional Results

Table 2: Accuracy and Win Rate: Existing Pretrained-Model

Model	Chain	Tree	Graph
Accuracy (%)			
Large Reasoning SOTA	82.86	78.19	76.75
Qwen2VL-7B	69.32	64.82	67.18
Win Rate (%)			
Large Reasoning SOTA	35.86	32.58	31.54
Qwen2VL-7B	28.96	23.88	26.06

Table 3: Win Rate and Accuracy Comparison: Topological Tuning

Model	**Chain**	**Tree**	**Graph**
Base Model			
Win Rate	43.14%	27.45%	29.41%
Overall Accuracy	61.00%	57.00%	57.00%
Topological Tuned Model			
Win Rate	45.16%	25.80%	29.03%
Overall Accuracy	89.66%	71.73%	75.47%

Table 4: Ablation study

Method	Accuracy	Chain	Tree	Graph
Baseline	79%	82%	55%	59%
Chain-only fine-tuned	81.97%	83.73%	61.23%	74.20%
Multi-topo (ours)	84.15%	86.33%	53.33%	78.95%