Optimizing Chain-of-Thought Reasoning through Preference-Based Learning

Abstract

Chain-of-Thought (CoT) prompting has significantly advanced the complex reasoning capabilities of Large Language Models (LLMs) by enabling them to generate explicit, step-by-step reasoning paths. However, this approach is often constrained by high inference complexity, the need for high-quality reasoning data, and challenges in transferring these capabilities to smaller models or new modalities. This paper synthesizes recent advancements in optimizing CoT reasoning through preference-based learning frameworks, particularly Direct Preference Optimization (DPO) and its variants. We examine several novel methodologies, including Chain of Preference Optimization (CPO) for improving efficiency, self-training with DPO for enhancing smaller models, and techniques for adapting CoT to multimodal and specialized domains like Vision Language Models (VLMs) and Text-to-SQL. The collective findings demonstrate that preference optimization provides a robust and versatile mechanism for refining reasoning processes. By learning to discern and favor superior reasoning chains, models achieve significant performance gains, enhanced efficiency, and improved applicability across a diverse range of complex tasks.

---

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable abilities in a wide array of natural language tasks. A key breakthrough in unlocking their potential for complex problem-solving was the development of Chain-of-Thought (CoT) reasoning [1]. By instructing models to "think step by step," CoT elicits an intermediate reasoning process that breaks down a complex problem into manageable parts, often leading to more accurate and interpretable solutions.

Despite its success, the standard CoT paradigm faces several significant challenges. First, advanced reasoning strategies that explore multiple paths, such as Tree-of-Thought (ToT), offer superior performance but at the cost of a substantial increase in inference complexity and computational resources [1]. Second, the efficacy of CoT fine-tuning is heavily dependent on the availability of high-quality datasets rich with explicit rationales. Many existing datasets, particularly in specialized domains like Text-to-SQL or multimodal reasoning, contain only final answers, which is insufficient for training robust reasoning capabilities [3, 5]. Finally, transferring the powerful reasoning abilities of state-of-the-art models to smaller, more efficient models remains a costly and often unstable process, typically relying on knowledge distillation from proprietary systems [2].

To address these limitations, a new paradigm centered on preference-based learning has emerged. Techniques like Direct Preference Optimization (DPO) offer a more direct and stable way to align models with human or machine-generated preferences. Instead of relying solely on ground-truth demonstrations, these methods train models by presenting them with pairs of responses—one preferred and one dis-preferred—and optimizing the

model to increase the likelihood of generating the preferred output.

This paper provides a comprehensive overview of recent research that leverages preference optimization to enhance and refine CoT reasoning. We synthesize findings from several key studies that apply these techniques across different contexts:
1. Improving Efficiency: Using Chain of Preference Optimization (CPO) to achieve the performance of complex search-based methods like ToT with the efficiency of standard CoT [1].
2. Enhancing Smaller Models: Employing self-training with DPO to improve the reasoning abilities of small-scale LMs without relying on expensive distillation from larger models [2].
3. Adapting to Multimodal Domains: Applying DPO to Vision Language Models (VLMs) through both supervised data enrichment and novel unsupervised frameworks [3, 4].
4. Unlocking Domain-Specific Reasoning: Demonstrating the critical need for CoT-augmented data to successfully apply DPO in specialized domains like Text-to-SQL [5].

By examining these diverse applications, we illustrate how preference-based learning is becoming a cornerstone for building more capable, efficient, and versatile reasoning systems.

---

2. Related Work

2.1. Chain-of-Thought Reasoning and its Variants

Chain-of-Thought (CoT) reasoning was a seminal development that prompted LLMs to produce intermediate steps before arriving at a final answer, significantly improving performance on arithmetic, commonsense, and symbolic reasoning tasks. This inspired more advanced reasoning structures, most notably Tree-of-Thought (ToT), which allows models to explore multiple reasoning paths in a tree-like structure. While ToT often yields superior results by enabling self-evaluation and backtracking, its search process introduces a significant computational overhead during inference [1].

2.2. Preference-Based Fine-Tuning

Aligning LLMs with desired behaviors has traditionally been accomplished through methods like Reinforcement Learning from Human Feedback (RLHF). However, RLHF involves training a separate reward model and can be complex and unstable. Direct Preference Optimization (DPO) emerged as a more direct and efficient alternative. DPO reframes the alignment problem as a simple classification loss on pairs of preferred and rejected responses, directly optimizing the language model policy without an explicit reward model. Its stability and effectiveness have made it a popular choice for fine-tuning models on complex tasks [2, 5].

2.3. Knowledge Distillation and Self-Training

Improving the capabilities of smaller, more resource-efficient LMs is a crucial area of research. A common approach is knowledge distillation, where a smaller "student" model is trained on the outputs of a larger, more powerful "teacher" model (e.g., GPT-4). While effective, this can be costly and dependent on proprietary APIs. An alternative is self-training, where a model learns from its own generated outputs. This approach promotes self-improvement and can be made more effective by incorporating mechanisms to filter or rank the model's own generations, a process that aligns naturally with preference optimization [2].

---

3. Methodology

The core methodology across the reviewed literature is the application of preference optimization to refine the step-by-step generation process inherent in CoT reasoning. This section synthesizes the distinct approaches used to achieve this goal across various models and domains.

3.1. Chain of Preference Optimization (CPO)

To bridge the performance gap between CoT and ToT without sacrificing efficiency, Zhang et al. (2024) proposed Chain of Preference Optimization (CPO). This method leverages the rich information generated during a ToT tree-search process to create preference data. Specifically, each step along the reasoning paths in the ToT search tree is compared and ranked. CPO then fine-tunes a base LLM by aligning its step-by-step CoT generations with the preferred intermediate steps derived from ToT. This allows the model to internalize the superior decision-making of the search-based process while maintaining the simple, sequential generation of standard CoT at inference time [1].

3.2. Self-Training with DPO for Small-Scale LMs

To enhance the reasoning of smaller models cost-effectively, Wang et al. (2024) developed a self-training framework augmented with DPO. The process begins with a small-scale LM generating multiple reasoning chains for a given problem. These outputs are then used to create preference pairs (`preferred`, `rejected`). The DPO algorithm is subsequently applied to fine-tune the model on this self-generated data, guiding it to produce more accurate and diverse reasoning. This creates a self-improvement loop where the model learns to refine its own reasoning abilities without relying on external, more powerful teacher models [2].

3.3. CoT Optimization in Vision Language Models (VLMs)

Adapting CoT reasoning to the multimodal domain presents unique challenges, primarily due to the scarcity of training data containing detailed visual rationales. Two distinct preference-based approaches were explored.

* Supervised Data Enrichment with DPO: Zhang et al. (2024) employed a two-stage approach. First, they enriched existing VLM datasets by using a highly capable model (GPT-4o) to generate detailed rationales for image-based questions. This created a high-quality CoT dataset. Second, they used DPO to fine-tune the

VLM, constructing preference pairs from correct (positive) and incorrect (negative) model-generated reasoning chains to further refine its reasoning quality [3].

* Unsupervised Visual CoT (UV-CoT): Addressing the need for labeled data, Zhao et al. (2025) proposed an unsupervised framework. UV-CoT uses a target MLLM to generate responses, including bounding boxes identifying relevant image regions. A separate, more powerful evaluator MLLM then ranks these responses based on the plausibility of the generated regions. These machine-generated rankings serve as preference signals to train the target model via DPO, creating a fully unsupervised pipeline that eliminates the need for human annotations [4].

3.4. Augmenting Domain-Specific Data for DPO

In specialized domains like Text-to-SQL, datasets often lack intermediate reasoning steps, only providing the final "gold" SQL query. Liu et al. (2025) discovered that applying DPO directly in such contexts often fails or even degrades performance. Their key insight was that DPO requires a structured reasoning process to optimize effectively. Their methodology involves first augmenting the Text-to-SQL dataset by synthetically generating CoT-style reasoning paths that lead to the final query. Only after this data enrichment step is DPO applied, using the gold query path as the preferred response and model-generated incorrect paths as the rejected ones. This provides the necessary structure for DPO to learn the correct reasoning process [5].

---

4. Experiments

To validate these methodologies, the researchers conducted a series of experiments across diverse datasets, models, and tasks, consistently comparing their proposed methods against strong baselines.

* Datasets: The experiments utilized a wide range of benchmarks, including standard reasoning datasets for question answering, fact verification, and arithmetic reasoning [1]; various mathematical reasoning tasks [2]; benchmark VLM datasets [3]; and specialized corpora like Text-to-SQL [5]. The UV-CoT study also included six main datasets and four unseen datasets to test for zero-shot generalization [4].

* Models: The studies involved various model types, demonstrating the broad applicability of preference optimization. These included general-purpose LLMs, small-scale LMs, and Vision Language Models (VLMs) such as LLaVA-1.5-7B and OmniLLM-12B [1, 2, 3, 4].

* Baselines: The proposed methods were rigorously evaluated against relevant state-of-the-art and standard approaches. Baselines included:
* Standard Chain-of-Thought (CoT) decoding and the more complex Tree-of-Thought (ToT) [1].
* Knowledge distillation from large proprietary LMs and conventional self-training methods [2].
* VLMs trained on datasets with minimal rationales and standard Supervised Fine-Tuning (SFT) with labeled data [3, 4].
* DPO applied directly to datasets without CoT augmentation [5].

* Evaluation: Performance was measured by comparing the outputs of the optimized models against the baselines on the respective benchmark tasks. The primary goal was to assess improvements in reasoning

accuracy, task performance, and, in some cases, generalization to unseen data.

---

5. Results

The experimental results across all studies consistently demonstrated the effectiveness of using preference optimization to improve CoT reasoning.

* CPO Achieves ToT Performance with CoT Efficiency: The CPO-trained model significantly improved LLM performance on complex problem-solving tasks. It successfully achieved results comparable or superior to the powerful but computationally expensive ToT baseline, while retaining the single-pass inference efficiency of standard CoT [1].

* Self-Training with DPO Enhances Small LMs: The self-training framework with DPO was shown to be a cost-effective and scalable solution for improving the reasoning abilities of small-scale LMs. This method outperformed conventional self-training and provided a viable alternative to costly knowledge distillation from proprietary models [2].

* Preference Optimization Boosts VLM Reasoning: In the multimodal domain, both supervised and unsupervised approaches yielded significant gains.
* Enriching data with distilled rationales and applying DPO led to substantial improvements in VLM CoT reasoning on benchmark datasets and better generalization to direct-answer tasks [3].
* The unsupervised UV-CoT framework surpassed state-of-the-art textual and visual CoT methods, including fully supervised baselines, and demonstrated strong zero-shot generalization capabilities on unseen datasets [4].

* CoT is a Prerequisite for Effective DPO in Text-to-SQL: The study by Liu et al. (2025) provided a critical insight: DPO's potential is only unlocked when applied to structured reasoning paths. Augmenting Text-to-SQL datasets with synthetic CoT solutions led to consistent and significant performance improvements for the first time in this domain. Ablation studies confirmed that the presence of a CoT structure mitigates reward hacking and strengthens the model's discriminative capabilities [5].

---

6. Discussion

The collective results from these studies offer several key insights into the optimization of CoT reasoning.

First, preference signals are a powerful tool for refining reasoning. Whether these preferences are derived from a complex search algorithm (ToT), a separate evaluator model, or self-generated pairs, they provide a targeted training signal that guides the model to understand why one reasoning path is better than another. This is a more nuanced form of learning than simply mimicking a single ground-truth solution, as is done in standard supervised fine-tuning.

Second, the quality and structure of training data are paramount. The research in both VLMs and Text-to-SQL highlights that a primary bottleneck for advanced reasoning is the lack of datasets with explicit rationales [3, 5]. The success of methods that synthetically generate or distill these rationales underscores a critical principle: to learn reasoning, a model must be trained on reasoning. This finding suggests that future efforts in dataset creation should prioritize the inclusion of step-by-step thought processes.

Third, these new methods offer pathways to greater efficiency and scalability. CPO provides a way to "bake in" the benefits of complex search methods into an efficient, sequential model [1]. Similarly, self-training with DPO democratizes reasoning improvements for smaller models, reducing reliance on massive, centralized systems [2]. The success of unsupervised VLM training further points towards a future where high-quality reasoning models can be developed without prohibitively expensive human annotation [4].

Finally, there is a clear synergy between CoT and DPO. The Text-to-SQL study reveals that DPO is not a universal solution that can be applied naively. It excels at optimizing a given policy over a distribution of possible outputs. CoT provides the necessary structure—a sequence of thoughts—over which this optimization can be effectively performed. Without this structure, the model may resort to "reward hacking" by focusing on superficial features of the final answer rather than learning a robust reasoning process [5].

---

7. Conclusion

Chain-of-Thought reasoning has fundamentally changed how we approach complex problem-solving with LLMs. However, realizing its full potential requires moving beyond standard prompting and fine-tuning. The research synthesized in this paper demonstrates that preference-based learning, particularly using Direct Preference Optimization, offers a versatile and powerful framework for overcoming the key limitations of CoT.

These methodologies have proven effective at enhancing inference efficiency, scaling down reasoning capabilities to smaller models, and extending CoT to new and challenging domains like multimodal and Text-to-SQL reasoning. A recurring theme is the critical importance of structured, high-quality reasoning data, which can be obtained through distillation, synthetic generation, or unsupervised preference signals. The synergistic relationship between the explicit structure of CoT and the refining power of DPO points toward a new frontier in developing more reliable, efficient, and broadly applicable AI reasoning systems.

---

8. Future Work

Based on the findings discussed, several promising directions for future research emerge:

* Generalizing Unsupervised Preference Learning: The success of UV-CoT in the visual domain invites exploration into unsupervised or weakly-supervised preference learning frameworks for other modalities and complex tasks where labeled reasoning data is scarce.

* Advanced Synthetic Data Generation: Further research is needed to develop more sophisticated and reliable methods for automatically generating high-quality CoT data for any domain, extending the work done for Text-to-SQL to other structured prediction tasks.
* Hybrid Training Methodologies: Investigating hybrid models that combine different techniques—for example, using an unsupervised evaluator to provide preference signals for a self-training loop—could lead to fully autonomous, self-improving reasoning agents.
* Theoretical Foundations: A deeper theoretical investigation into the interplay between CoT and DPO is warranted. Understanding precisely why explicit reasoning steps are crucial for mitigating reward hacking and enabling effective preference optimization could inform the design of future alignment techniques for a wide range of complex tasks.

References

[1] Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, Min Lin. (2024). Chain of Preference Optimization: Improving Chain-of-Thought Reasoning in LLMs.

[2] Tianduo Wang, Shichen Li, Wei Lu. (2024). Self-Training with Direct Preference Optimization Improves Chain-of-Thought Reasoning.

[3] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, Yiming Yang. (2024). Improve Vision Language Model Chain-of-thought Reasoning.

[4] Kesen Zhao, Beier Zhu, Qianru Sun, H. Zhang. (2025). Unsupervised Visual Chain-of-Thought Reasoning via Preference Optimization.

[5] Hanbing Liu, Haoyang Li, Xiaokang Zhang, Ruotong Chen, Haiyong Xu, Tian Tian, Qi Qi, Jing Zhang. (2025). Uncovering the Impact of Chain-of-Thought Reasoning for Direct Preference Optimization: Lessons from Text-to-SQL.