CoT(Chain of Thought) as Intended, Not Emergent, Through Mathematical Problem Solving

Anonymous CVPR submission

Paper ID

Abstract

Until recently, Chain of Thought (CoT) reasoning has been regarded as a core capability of large language models (LLMs) for solving complex reasoning tasks. It has also been considered an emergent ability that arises from the sheer scale of such models. However, this perspective has posed limitations when applying CoT to small language models (SLMs), which are gaining attention for their efficiency and security advantages.

In this study, we argue that CoT, previously viewed as an emergent property of LLMs, can instead be intentionally trained in SLMs for specific tasks. Through empirical experiments, we demonstrate that CoT is not unique to LLMs; smaller models are also capable of exhibiting effective reasoning performance when appropriately fine-tuned and equipped with targeted techniques.

These findings suggest that CoT should not be exclusively categorized as an emergent ability of large-scale models, but rather as an intended ability that can be systematically induced in smaller models. This opens up the possibility for SLMs to achieve reasoning capabilities comparable to LLMs, broadening the scope of their practical applications.

1. Introduction

Large language models (LLMs) have demonstrated unprecedented performance in natural language processing and brought transformative impacts across various industries. In particular, the emergence of Chain-of-Thought (CoT) prompting has been a significant development, showing that LLMs can perform complex, multi-step reasoning beyond mere text generation. This advancement marks a shift in AI from retrieving predefined knowledge to engaging in logical reasoning to derive novel conclusions. However, CoT abilities have been regarded as an *emergent property* that appears only in models with hundreds of billions of parameters, imposing clear limitations due to the enormous computational resources, operational costs, and environmental burden required by such large-scale models.

Amid this context, the industry trend is shifting from "larger models" toward "more efficient models." There is increasing demand for small language models (SLMs) specialized for specific tasks, motivated by cost efficiency, faster response times, and strong data security without relying on API calls. SLMs have the potential to enable AI services in on-device or edge computing environments, free from many constraints associated with LLMs. Nevertheless, conventional wisdom holds that SLMs face a *reasoning gap*: their limited scale makes it difficult to exhibit high-level reasoning abilities such as CoT, restricting their applicability in domains that require complex problem-solving.

In this study, we challenge this limitation by proposing a new hypothesis: **CoT is not an emergent property of large models, but an *intended ability* that can be deliberately instilled in small models through purpose-driven training**. In other words, reasoning skills do not need to emerge naturally with model size; they can be explicitly taught through carefully designed data and learning strategies. To validate this, we developed an RLHF pipeline for the Korean language KoGPT2 model with 125 million parameters. By strategically combining general conversational data with mathematical CoT data, we fine-tuned the model and evaluated its performance both qualitatively and quantitatively, demonstrating the potential for reasoning ability in SLMs.

This work makes three main contributions. First, we present a concrete methodology for training CoT abilities in a publicly available Korean SLM using an RLHF pipeline. Second, we demonstrate the effectiveness of a data composition strategy (oversampling) that combines reasoning and conversational datasets. Finally, our results challenge the conventional notion that reasoning abilities are purely emergent and provide empirical evidence supporting the development of low-cost, high-efficiency SLMs for specialized domains.

The remainder of this paper is structured as follows. Section 2 reviews related work on CoT, SLMs, and RLHF. Section 3 details the dataset composition, training pipeline, and evaluation methods used in this study. Section 4 presents experimental results and training analysis, while Section 5 discusses the implications and limitations. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Work

2.1. Chain-of-Thought (CoT) and Its Development

First introduced by Wei et al. (2022), Chain-of-Thought (CoT) is a prompting technique that guides large language models (LLMs) to generate step-by-step reasoning before producing a final answer. By mimicking the way humans solve complex problems through intermediate steps, CoT significantly improved LLM performance across arithmetic, commonsense, and symbolic reasoning tasks. Early studies, however, regarded this ability as an **emergent property** that only appeared in models with more than 100 billion parameters, framing model scale as the key prerequisite for reasoning capabilities.

Research on CoT has since advanced in various directions. Kojima et al. (2022) proposed Zero-shot-CoT, showing that even a simple instruction such as *"Let's think step by step"* can elicit reasoning chains, suggesting that reasoning abilities are inherently embedded within LLMs. Furthermore, Wang et al. (2022) introduced the Self-Consistency method, where multiple reasoning paths are generated and the most consistent answer is selected via majority voting, greatly enhancing accuracy and robustness. While these studies expanded the potential of CoT, the discussion remained centered on large-scale models.

2.2. The Rise of Small Language Models (SLMs) and High-Quality Data

Small language models (SLMs), typically containing millions to billions of parameters, operate with far fewer computational resources than LLMs. Initially dismissed as scaled-down versions of LLMs with clear performance limitations, SLMs have recently gained renewed attention. Microsoft's Phi series (Gunasekar et al., 2023) demonstrated that training on *textbook-quality* data could enable smaller models to outperform much larger counterparts, highlighting a paradigm shift: **data quality matters more than sheer scale**.

Similarly, recent SLMs such as Mistral 7B and Google's Gemma have achieved LLM-comparable performance in specialized domains through carefully curated data and efficient architectures. This shift underpins the theoretical foundation of our work, which hypothesizes that complex reasoning can be intentionally taught to SLMs through high-quality training data. If reasoning quality is indeed determined by data, then providing SLMs with reasoning-rich supervision should directly impart reasoning abilities.

2.3. Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF) aligns language model outputs with human values and preferences, and was popularized through OpenAI's InstructGPT (Ouyang et al., 2022). The process typically involves three stages: (1) **Supervised Fine-Tuning (SFT)**, where the model learns from human demonstrations; (2) training a **Reward Model (RM)** that scores outputs according to human preference; and (3) optimizing the model policy with reinforcement learning algorithms such as PPO, guided by the reward model.

RLHF has become the standard paradigm, enabling models to produce responses that are not only factually correct but also more useful, safe, and instruction-following. In this study, we adopt the RLHF framework under the assumption that it is especially effective for training abilities like CoT, where the *quality of reasoning steps* matters as much as the final answer.

2.4. Transferring Abilities to Smaller Models

Research on transferring LLM capabilities to SLMs has largely focused on knowledge distillation. In this approach, a smaller *student model* is trained to mimic the outputs or intermediate representations of a larger *teacher model*. More recently, distillation has been applied to reasoning, where detailed CoT solutions generated by powerful models such as GPT-4 are used as SFT data for training smaller models. This method shows the feasibility of directly imparting reasoning skills through explicit supervision.

Building on this line of work, our study goes a step further by empirically demonstrating the induction of reasoning ability in publicly available Korean SLMs using an RLHF pipeline. We provide quantitative evaluations to validate this approach, distinguishing our work as one of the first practical implementations of intentional CoT training in small-scale models.

3. Methodology

This study systematically designs experiments to intentionally train Chain-of-Thought (CoT) reasoning in small language models (SLMs), covering dataset preparation, model training, and evaluation. The base model used in our experiments is **skt/kogpt2-base-v2** with 125 million parameters.

3.1. Dataset Preparation and Pipeline Design

The core objective of this study is to verify whether direct fine-tuning using high-quality reasoning data can enhance the reasoning capabilities of SLMs. To this end, we set up a controlled experiment with baseline and experimental pipelines using an RLHF framework.

Baseline Pipeline: We utilized 12,000 SFT samples from KoChatGPT, consisting of general Korean question-answer pairs, to perform SFT, reward model (RM) training, and PPO. This serves as a reference point to measure the model's general performance.

Experimental Pipeline:

- **Initial Attempt (English Datasets):** Initially, 8,900 high-quality English math CoT samples from GSM8K were combined with the SFT dataset. However, the KoGPT2 tokenizer, optimized for Korean, treated most English words as unknown tokens, resulting in poor learning due to language mismatch.

- **Final Design (Korean Dataset):** To address this, we collected a large-scale Korean math CoT dataset of 890,000 samples (ChuGyouk/AI-MO-NuminaMath-CoT-Ko). For efficiency while maintaining data quality, 12,000 samples were randomly **undersampled** and combined with 12,000 SFT samples, resulting in a total of 24,000 SFT samples for the final dataset.

3.2. Role of High-Quality Reasoning Data

The Korean math CoT dataset used in this study contains not only problem statements and answers but also step-by-step reasoning processes leading to the correct solution. This **high-quality reasoning data** enables the model to learn the *patterns of reasoning* rather than merely memorizing answers, serving as a crucial factor in developing CoT capability.

3.3. Direct Fine-Tuning via RLHF

We consider the three-stage RLHF pipeline as a direct fine-tuning procedure to instill reasoning ability.

1. **Supervised Fine-Tuning (SFT):** Using the curated 24,000-sample dataset, the model learns the syntactic structure of CoT responses.

2. **Reward Modeling (RM):** A pre-built RM dataset is used to train a "judge" model capable of assessing which responses are logically superior.

3. **PPO Reinforcement Learning:** The trained RM is used as a reward function to optimize the SFT model via PPO. This stage refines the model's policy, encouraging it to generate high-quality reasoning steps that are both logical and closer to correct answers, beyond merely following CoT format.

4. Experimental Results and Analysis

Contrary to our initial hypothesis, training the KoGPT2 model to acquire CoT reasoning proved to be significantly challenging. In this section, we analyze the observations during training and qualitatively examine the outputs of the final models.

4.1. Training Analysis: Stable SFT vs. Unstable PPO

During the first stage of the experiment, supervised fine-tuning (SFT), training loss exhibited a stable decreasing trend. Figure 1 shows the training loss curve of the experimental SFT model trained on CoT data. As the figure indicates, loss steadily declined over time, suggesting that the model successfully learned the statistical patterns of the dataset—i.e., the syntactic structure of CoT responses.

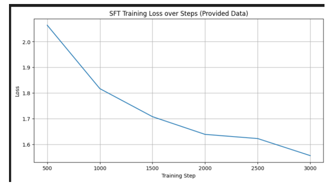

Figure 1. **(Figure 1: Training loss of the experimental SFT model)**

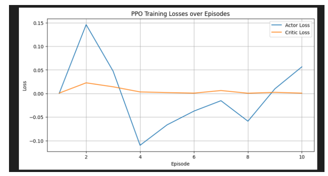

Figure 2. **(Figure 2: Actor and Critic losses during PPO training)**

However, instability arose during the PPO reinforcement learning stage. Figure 2 illustrates the Actor and Critic losses during PPO training. While the Critic loss remained near zero and stable, the Actor loss exhibited extreme volatility, with sudden spikes and drops as training episodes progressed. Notably, in the fourth episode, the Actor loss fell sharply below -0.1, indicating high instability.

Such instability in the Actor loss indicates that the model failed to learn a consistent policy. This may be due to the reward model (RM) inconsistently evaluating the logical quality of CoT responses, providing inaccurate reward signals, or the Actor model engaging in *reward hacking*—repeatedly generating meaningless text to maximize reward scores. Ultimately, despite apparent learning of CoT format during SFT, the PPO stage failed to internalize actual reasoning capabilities.

4.2. Qualitative Evaluation: Format Imitation vs. Logical Failure

A comparison between the PPO-trained experimental model (CoT applied) and the baseline model revealed that, while the experimental model successfully **imitated the CoT "format"**, it completely failed to internalize **logical reasoning**. Table 1 highlights major failure cases.

The analysis indicates that the experimental model successfully generated the phrase *"Let's think step by step"* and attempted stepwise reasoning. However, the content frequently deviated from the question's intent, contained factual errors, or consisted of unrelated mathematical symbols. This demonstrates that the model learned the **surface "style"** of CoT data but failed to comprehend the underlying **logical "meaning"**, highlighting a clear limitation in instilling actual reasoning capability in the SLM.

The analysis indicates that the experimental model successfully generated the phrase *"Let's think step by step"* and attempted stepwise reasoning. However, the content frequently deviated from the question's intent, contained factual errors, or consisted of unrelated mathematical symbols. This demonstrates that the model learned the **surface "style"** of CoT data but failed to comprehend the underlying **logical "meaning"**, highlighting a clear limitation in instilling actual reasoning capability in the SLM.

5. Discussion

The experiments in this study did not successfully confirm the initial hypothesis that CoT reasoning can be intentionally trained in SLMs. However, these failures provide valuable insights and highlight critical challenges for training reasoning capabilities in small models.

First, there are fundamental limitations of the base model. KoGPT2, with 125 million parameters, likely lacks the capacity to understand complex mathematical relationships and perform multi-step reasoning. This limitation may stem from the model architecture and size, which cannot be fully compensated for by high-quality data alone.

Second, there are challenges in reward modeling for reasoning ability. Current reward models are primarily effective in assessing fluency or formal completeness of sentences. Accurately evaluating the abstract concept of "logical correctness" in mathematical problem-solving and providing consistent reward signals remains extremely difficult. The instability observed during PPO training supports the possibility that inaccurate reward signals interfered with the Actor model's learning.

Third, there is a gap between format imitation and actual reasoning. The results indicate that SLMs can imitate the CoT format relatively easily, but understanding and executing the logical connections between steps is a fundamentally different task. This suggests that future research on SLM reasoning should explore new training and evaluation methods that verify not only the output form but also the validity of the reasoning process.

In conclusion, the RLHF pipeline applied in this study was insufficient to transfer complex reasoning abilities to the current SLM. This raises the fundamental question of whether reasoning capability remains an emergent property that only manifests in models above a certain scale.

6. Conclusion

This study aimed to train Chain-of-Thought (CoT) reasoning in the small language model KoGPT2 using an RLHF pipeline for solving mathematical problems. The results show that while the model fine-tuned with CoT data successfully learned the CoT format during the SFT stage, it exhibited instability during PPO reinforcement learning and ultimately failed to internalize logical reasoning capabilities.

These findings experimentally demonstrate the limitations of current standard RLHF methodologies in significantly enhancing the intrinsic reasoning ability of SLMs. The outcome is influenced by a combination of factors, including base model size, the sophistication of reward modeling, and the quality and quantity of training data.

Although the initial hypothesis was not confirmed, this study contributes by implementing a concrete pipeline for training CoT ability in SLMs and empirically presenting the key challenges encountered. Future work should explore using larger SLMs or novel approaches, such as **process-based reward models**, which supervise the validity of the reasoning process directly, to overcome the limitations identified in this study.

7. References

- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *arXiv preprint arXiv:2205.11916*.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171*.
- Wang, Z., Jiang, J., Qiu, T., Liu, H., Tang, X., & Yao, H. (2025). Efficient Long CoT Reasoning in Small Language Models. arXiv preprint arXiv:2505.18440.
- Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). Automatic Chain of Thought Prompting in Large Language Models. arXiv preprint arXiv:2210.03493.