OPENENDEDREASONER: SELF-IMPROVING OPEN-ENDED REASONING LLMS

Alkın Ünlü
alkinunl@gmail.com

ABSTRACT

In recent months, OpenAI o1 has shown promising progress in solving complex reasoning tasks by synthesizing chain-of-thoughts (CoT) before giving a final answer. This approach has demonstrated the potential to enhance performance on reasoning and coding tasks. However, existing methodologies remain limited by the need for human intervention, curated datasets, or grounded verifiers, especially the need for groundable topics such as mathematics, but a reinforcement learning (RL) approach has yet to be fully explored with open-ended reasoning tasks.

This paper introduces ***OpenEndedReasoner***, a novel framework for enabling fully open-ended self-improvement in reasoning LLMs. By leveraging the diversity of data at both pretraining and post-training stages, ***OpenEndedReasoner*** iteratively generates and filters high-quality reasoning traces without requiring human intervention or seed data. The framework consists of two key components: **Question Synthesis**, which synthesizes diverse and steerable prompts, and **Response Synthesis**, which leverages a reward model that evaluates reasoning quality and selects optimal responses. Through these stages, the model refines its reasoning capabilities, producing reasoning-heavy data to improve itself continuously.

We demonstrate that ***OpenEndedReasoner*** achieves substantial improvements in reasoning tasks, validated by a 17.68-point gain on the GPQA benchmark across four iterations. This framework highlights the feasibility of building open-ended, self-improving reasoning systems capable of evolving their reasoning abilities.

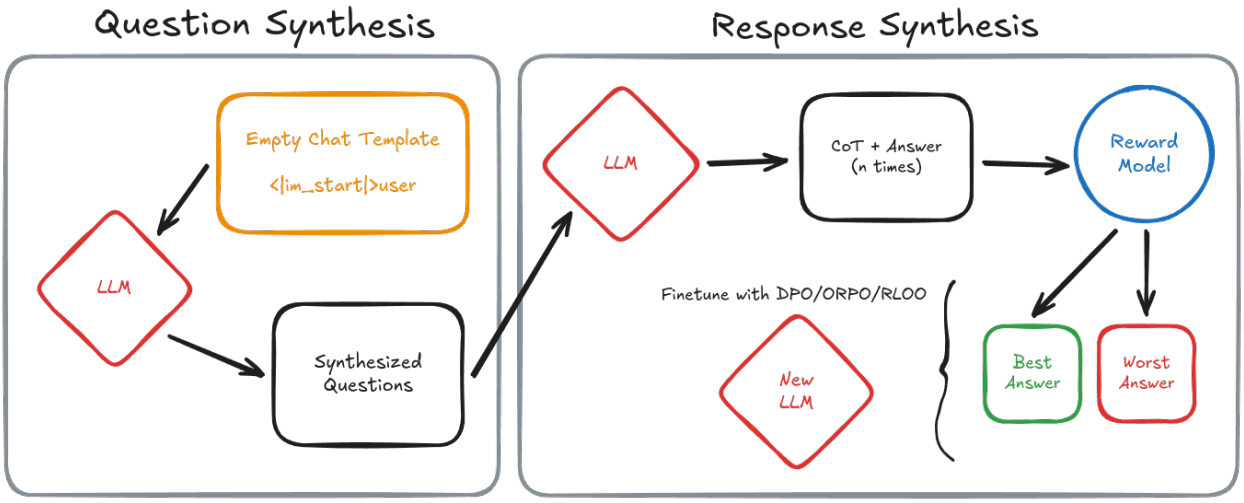

Figure 1: A simplified view of ***OpenEndedReasoner***. While other approaches require grounded data, our approach doesnt require any human intervention or seed data to improve its reasoning abilities.

1 Introduction

Recently, OpenAI announced o1, a large language model (LLM) that synthesizes long internal thoughts before giving a final answer. This is proven to be really effective in reasoning and coding tasks. This release also opened a new door to increase test-time compute. By bootstrapping already knowladgable LLMs with this ability, we can build systems that are closer to human thinking while still doing system-1 thinking ultimately. To achieve this, simply finetuning the base model with supervised-finetuning (SFT) isnt enough as the model has to have a deeper understanding internal thought structure. As a result, reinforcement learning (RL) is crucial to bootstrap this behavior to any LLM. However, RL has some limitations, such as the need for grounded verifiers, which cannot be done with open-ended tasks. In this paper we propose ***OpenEndedReasoner***, a fully open-ended framework that leverages the diversity of data in the pretraining and post-training stages of LLMs to synthesize steerable and diverse queries to then be answered multiple times and chosen with a reward model. We posit that the reward model will not be a bottleneck, as the CoT will also inputted, allowing the reward model to interpret the steps proposed by the model.

In this paper, we propose ***OpenEndedReasoner***, a framework that enables LLMs to self-improve by generating and filtering high-quality reasoning traces. ***OpenEndedReasoner*** presents a fully open-ended approach that consists of two main stages: **Question Synthesis** and **Response Synthesis**. A reward model evaluates and selects the best questions and responses, which are then used to fine-tune the model iteratively. Our approach shows significant gains in the GPQA benchmark, without a single human intervention or seed data, highlighting the potential of open-ended systems to generate high-quality, diverse datasets to self-improve reasoning.

Our main contributions are as follows:

- We propose a two-stage, reward-driven framework for question and response synthesis, enabling LLMs to self-improve and refine their own reasoning ability.
- We demonstrate significant improvements in GPQA scores across multiple iterations, validating the effectiveness of self-improvement in reasoning.
- We provide a comprehensive analysis of ***OpenEndedReasoner***'s question-answer generation capabilities, showing that self-synthesized data can lead to robust and nuanced improvements without external human supervision.

2 Related Work

The field of self-improving language models has seen significant advancements in recent years. A number of frameworks have been proposed to enhance the capabilities of large language models (LLMs) through self-generated data, self-play, self-reward and so on. However, these frameworks are limited in their ability to be both: fully open-ended and steerable.

Self-Rewarding Language Models Recent work has shown that LLMs can be used to self-critique and generate high-quality, diverse finetuning data for themselves. While these approaches have shown promising results, they lack the ability to be steerable, grounded and performant.

Self-Play Recent approaches also have been proposed to improve the capabilities of LLMs with the help of a self-play mechanism to strictly adhere to the inputted dataset. While these approaches have shown promising results, they are limited to the human-generated data that is inputted to them.

OpenEndedReasoner falls under the category of fully open-ended but steerable systems that can recursively improve reasoning, refine and synthesize their own training data.

3 Methodology

3.1 System Overview

The ***OpenEndedReasoner*** framework consists of two main components: **Question Synthesis** and **Response Synthesis**. The **Question Synthesis** stage generates diverse and steerable questions, while the **Response Synthesis** stage leverages a reward model to select high-quality examples, enabling iterative refinement in reasoning. Figure 1 provides a schematic overview of the overall system.

3.2 Question Synthesis

The thia stage starts with an empty chat template as the input for the LLM (e.g., `<|im_start|>user`). With this template, the LLM synthesizes a random user instruction. We find that capable model can generate pretty promising questions at temperature of 1, however smaller and less capable models may require lower temperatures to generate coherent questions.

- **Prompt Initialization**: An empty template is used to prompt the LLM to generate questions.
- **Question Generation**: The model generates questions with temperature of 1.
- **Question Steerability**: We can set the system prompt to steer the generated questions (e.g., `The user has a question about reasoning, try your best to answer it.`).

To generate question with the LLM, we use the following template:

```
<|im_start|>system
You are a helpful assistant.
Try your best to answer the users reasoning question.<|im_end|>
<|im_start|>user
```

With these prompts we can generate quality, diverse and steerable questions in an open-ended fashion.

3.3 Response Synthesis

In this stage, we sample n number of answers with reasoning traces for each synthesized question. After sampling, we use our reward model to pick the highest-scoring and the lowest-scoring answer to generate a preference pair.

- **Answer Generation**: For each synthesized question, the model generates multiple candidate CoT and answers.
- **Reward-Based Selection**: The reward model scores each candidate CoT and answer, selecting the highest and lowest scoring answers. We find that the reward model is cannot be a bottleneck, as CoT will also be inputted with final answer, enabling our reward model to interpret the steps proposed by the model. This helps the reward model to follow the CoT and generate high-quality rewards.

4 Experimental Setup

4.1 Setup

With the base model being `Qwen2.5-1.5B-Instruct`, we sample 600 data points from the latest iteration of our model. We finetune the latest iteration of our model with Unsloth SFT.

4.2 Evaluation Metrics

The GPQA score evaluates the model's question-answering performance by measuring factors such as accuracy, relevance, and diversity of generated answers.

4.3 Baseline

We initialized ***OpenEndedReasoner*** with `Qwen2.5-1.5B-Instruct` as the baseline model. This model serves as the starting point for iterative self-improvement, with initial GPQA scores of 16.41.

5 Results

5.1 Iterative Improvement on GPQA

Table 1 shows the GPQA scores across five iterations, demonstrating consistent improvement with each iteration.

Iteration	GPQA Score
0	16.41
1	28.03
2	30.05
3	32.32
4	34.09

Table 1: GPQA scores across iterations, with 0 being the baseline.

6 Conclusion

We presented ***OpenEndedReasoner***, a novel framework for the open-ended self-improvement of LLMs through open-ended question-answer synthesis. Over four iterations, ***OpenEndedReasoner*** demonstrated a 17.68-point improvement on the GPQA benchmark with no signs of plateau, highlighting the future of self-synthesized data in driving model enhancement.

References

[1] Author A., Title of Paper A. Journal Name, Year.
[2] Author B., Title of Paper B. Journal Name, Year.
[3] Author C., Title of Paper C. Journal Name, Year.
[4] Author D., Title of Paper D. Journal Name, Year.