OpenR: An Open Source Framework for Advanced Reasoning with Large Language Models

Jun Wang[1], Meng Fang[2], Ziyu Wan[3], Muning Wen[3], Jiachen Zhu[3], Anjie Liu[4], Ziqin Gong[4],
Yan Song[1], Lei Chen[4], Lionel M. Ni[4], Linyi Yang[5], Ying Wen[3], Weinan Zhang[3]

[1] University College London [2] University of Liverpool [3] Shanghai Jiao Tong University
[4] The Hong Kong University of Science and Technology (Guangzhou) [5] Westlake University

Abstract

In this technical report, we introduce *OpenR*, an open-source framework designed to integrate key components for enhancing the reasoning capabilities of large language models (LLMs). *OpenR* unifies data acquisition, reinforcement learning training (both online and offline), and non-autoregressive decoding into a cohesive software platform. Our goal is to establish an open-source platform and community to accelerate the development of LLM reasoning. Inspired by the success of OpenAI's o1 model, which demonstrated improved reasoning abilities through step-by-step reasoning and reinforcement learning, *OpenR* integrates test-time compute, reinforcement learning, and process supervision to improve reasoning in LLMs. Our work is the first to provide an open-source framework that explores the core techniques of OpenAI's o1 model with reinforcement learning, achieving advanced reasoning capabilities beyond traditional autoregressive methods. We demonstrate the efficacy of *OpenR* by evaluating it on the MATH dataset, utilising publicly available data and search methods. Our initial experiments confirm substantial gains, with relative improvements in reasoning and performance driven by test-time computation and reinforcement learning through process reward models. The *OpenR* framework, including code, models, and datasets, is accessible at `https://openreasoner.github.io`.

1 Introduction

OpenAI has recently unveiled o1 [OpenAI, 2024], a groundbreaking large language model (LLM) that represents a giant leap forward in strong AI. The model is reported to be five times more proficient in math and coding compared to the previous GPT-4o, specifically displaying exceptional performance across various domains: it ranks in the 89th percentile for competitive programming, places among the top 500 students in a prestigious US math olympiad qualifier, and surpasses human PhD-level accuracy in physics, biology, and chemistry benchmarks. Trained using reinforcement learning techniques, o1 excels in complex reasoning tasks by explicitly embedding a *native* "Chain-of-Thought" (NCoT) process in LLMs, which allows it to "deep think" through step-by-step reasoning before generating responses. A key innovation of o1 is that it allows spending more time reasoning during the inference process, marking a shift from fast, direct responses to slow, deliberate, multi-step inference-time computation, as illustrated in Figure 1.

Interestingly, in human cognition, two correlated yet distinct modes of cognitive processing are presented to guide human decision-making and behaviours [Kahneman, 2011], each of which has the partial distinction between brain circuits and neural pathways. System 1 thinking is fast, automatic,

Correspondence to: Jun Wang and Meng Fang.

Technical Report. Work in progress.

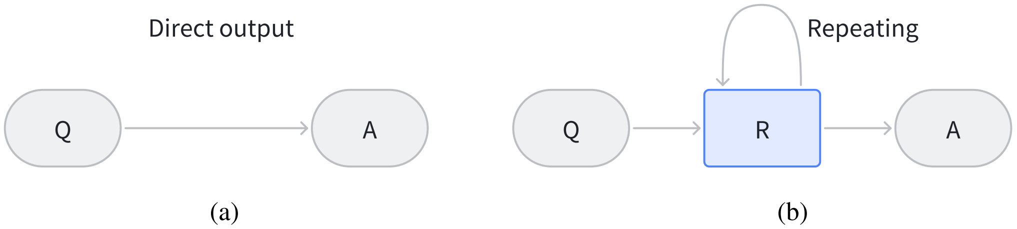

Figure 1: Inference-time computation. (a) An autoregressive LLM directly generates an answer (A) by conditioning the given question (Q). (b) The concept of chain of thought, or step-by-step thinking, involves incorporating intermediate reasoning steps (R) before arriving at the final answer (A). These repeated operations allow for 1) revisiting and revising prior outputs, 2) progressing to subsequent reasoning stages, and 3) exploring multiple reasoning paths or trajectories.

and intuitive, operating effortlessly and often unconsciously. It relies on neural pathways that enable rapid processing, especially in situations needing quick reactions or when cognitive resources are constrained. System 2 thinking is deliberate, effortful, and conscious, involving focused attention and analytical reasoning. It processes information more slowly and is used for complex problem-solving, logical reasoning, and decision-making tasks.

o1 model is an exciting development for AI, as LLMs can now not only generate rapid responses using learned patterns but, more significantly, simulate complex reasoning processes through mechanisms like chain-of-thought or other forms of search, similar to how humans engage in deeper, step-by-step thinking. o1's improved reasoning skills induce implications for multiple fields, including science, coding, and mathematics. In coding competitions, a specialised version of o1 achieved impressive results, scoring in the 49th percentile in the 2024 International Olympiad in Informatics and outperforming 93% of human competitors in simulated Codeforces contests. Beyond its technical capabilities, o1 also represents progress in AI safety and alignment. The model's chain of thought reasoning provides new opportunities for integrating human values and principles, resulting in improved performance on safety evaluations and jailbreak tests.

The idea of chain-of-thought reasoning [Wei et al., 2022] and step-by-step thinking in large language models (LLMs) is not new. Previous research has shown that simply adding instructions like "describe your reasoning in steps" or "explain your answer step by step" to the input questions or providing a few shot examples can trigger LLMs to generate intermediate reasoning steps (as illustrated in Figure 1) and subsequently improve problem-solving, especially in tasks like math and coding [Wei et al., 2022, Nye et al., 2021]. However, these approaches build on existing LLMs without truly embedding the chain of thought ability within the models themselves. As a result, LLMs cannot inherently learn this reasoning capability, leading to active research on how to integrate it directly into model training. Proposed methods range from collecting specialised training data to building reward models [Ouyang et al., 2022, Li et al., 2022, Luo et al., 2024] and increasing the computational complexity of decoding [Snell et al., 2024, Wu et al., 2024], but none have yet achieved significant performance breakthroughs at scale.

It remains unclear whether o1's innovation is rooted in the model itself, rather than relying on external prompting systems. If it indeed involves explicitly embedding step-by-step reasoning natively within the architecture, this would represent a significant breakthrough. Building on substantial performance gains, o1 has shown that the scaling principles traditionally applied during training [Kaplan et al., 2020, Snell et al., 2024] are now relevant to the inference phase. We should reallocate our computational focus, balancing pre-training efforts with efficient use of inference-time computation. Allowing LLMs to enhance their outputs with increased test-time computing is an essential step towards creating generally self-improving agents capable of managing open-ended strong reasoning and decision-making tasks. This direction, which we refer to as LLM-Native Chain-of-Thought (NativeCoT), should be able to inherently mirror the deliberate, analytical process possessed by human's System 2 thinking [Kahneman, 2011].

In this report, we present *OpenR*, an open-source framework built on the principles behind OpenAI's o1 model, designed to replicate and extend its reasoning capabilities. Our approach focuses on improving LLM reasoning by integrating process supervision, reinforcement learning (RL), and inference-time computation strategies such as guided search. *OpenR* implements key components such as data augmentation for process supervision, policy learning via RL, and efficient decoding algorithms. By doing so, it shifts the focus from merely scaling model parameters during pre-training to leveraging smarter inference strategies at test time. These techniques help the model refine its reasoning step by step, allowing it to pause, evaluate intermediate reasoning, and select better solution pathways during test-time computation. Through experiments on publicly available benchmarks, such as the MATH dataset, we show that the combination of process reward models and guided search improves test-time reasoning performance by approximately 10%.

In summary, we introduce *OpenR*, an open-source framework that integrates test-time computation and process supervision to enhance reasoning in LLMs, providing an open platform with models, data, and code to foster collaboration and accelerate research in LLM reasoning. To our knowledge, *OpenR* is the first open-source framework to explore the core methods of OpenAI's o1 model with reinforcement learning techniques. The framework includes reinforcement learning algorithms designed to optimize decision-making during training, enabling more accurate and deliberate step-by-step reasoning. Additionally, *OpenR* provides tools for generating synthetic process reward data, reducing dependence on costly human annotations and supporting scalable process supervision. Through experiments, we demonstrate the effectiveness of process reward models and test-time guided search.

2 Related Work

Key references in the field of improving reasoning capabilities in large language models (LLMs) highlight several innovative approaches, including inference-time computing, process reward models, and data acquisition methods.

Inference-time Computing. To discuss the role of inference-time computation in large language models (LLMs), recent studies have focused on optimizing the efficiency and effectiveness of reasoning during the inference process rather than merely relying on the scaling law of training-time computing. A pivotal study, Feng et al. [2024] demonstrate the benefits of using MCTS as a decoding mechanism, which enhances inference computation by actively planning and selecting higher-quality responses. This approach aligns with the reasoning-as-planning approach proposed in Hao et al. [2023], where reasoning is viewed as a process similar to planning in decision-making processes, further underscoring the centrality of step-wise reasoning at inference time. In recent, the work [Snell et al., 2024] reinforces that optimizing inference strategies can yield superior performance gains compared to simply increasing model size, underscoring the critical role of test-time computation. Finally, this is complemented by the findings of work [Goyal et al., 2023], which introduces an implicit reasoning model by incorporating pause tokens to encourage deliberate reasoning during generation. Collectively, these recent advances suggest the growing recognition of inference-time optimisation – whether through planning-based reasoning models or computational optimisation – as a critical factor in improving LLM capabilities, advocating for strategies that enhance reasoning, planning, and compute efficiency beyond mere training-time scaling.

From Outcome Supervision to Process Supervision. The shift from Outcome Supervision to Process Supervision in language model training has gained prominence in recent research, particularly with respect to enhancing reasoning capabilities. The foundational work by Cobbe et al. [2021a] introduces Outcome-supervised Reward Models (ORM) and the widely used math reasoning dataset, GSM8K, where verifiers are trained to assess the final correctness of generated solutions. While ORM plays a crucial role in the early stage, it primarily focuses on evaluating the end result rather than the reasoning steps leading to the final output.

Building on this, the concept of process reward models (PRM) is introduced as a more granular and transparent approach. With both ORM and PRM, DeepMind proposes the idea of supervising intermediate reasoning steps alongside the final outcome, allowing for more detailed feedback during the reasoning process [Uesato et al., 2022]. This research laid the groundwork for subsequent developments in process-based verification. On the other hand, OpenAI's work [Lightman et al., 2023] continues this trend by refining PRM through a follow-up study that emphasizes verifying each intermediate step in reasoning tasks by providing a high-quality human-labelled process-supervision dataset, namely PRM800K, which has been enriched in our work.