Novelty Assessment Report

Paper: OpenThoughts: Data Recipes for Reasoning Models

PDF URL: https://openreview.net/pdf?id=7xjoTuaNmN

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

Reasoning models have made rapid progress on many benchmarks involving math, code, and science. Yet, there are still many open questions about the best train- ing recipes for reasoning since state-of-the-art models often rely on proprietary datasets with little to no public information available. To address this, the goal of the OpenThoughts project is to create open-source datasets for training reasoning models. Our OpenThoughts2-1M dataset led to OpenThinker2-32B, the first model trained on public reasoning data to match DeepSeek-R1-Distill-32B on standard reasoning benchmarks such as AIME and LiveCodeBench. We then improve our dataset further by systematically investigating each step of our data genera- tion pipeline with 1,000+ controlled experiments, which led to OpenThoughts3. Scaling the pipeline to 1.2M examples and using QwQ-32B as teacher yields our OpenThinker3-7B model, which achieves state-of-the-art results: 53% on AIME 2025, 51% on LiveCodeBench 06/24-01/25, and 54% on GPQA Dia- mond – improvements of 15.3, 17.2, and 20.5 percentage points compared to the DeepSeek-R1-Distill-Qwen-7B. All of our datasets and models are available on ANONYMIZED.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Creating Open-Source Datasets for Training Reasoning Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Dataset Construction Methodologies**
- **Domain-Specific Reasoning Datasets**
- **General Reasoning Datasets and Benchmarks**
- **Training Frameworks and Model Architectures**

Complete Taxonomy Tree

- Creating Open-Source Datasets for Training Reasoning Models Survey Taxonomy
- Dataset Construction Methodologies
  - Synthetic Data Generation (5 papers)
  - [9] Key-point-driven data synthesis with its enhancement on mathematical reasoning (Huang, 2025) View paper
  - [11] SynLogic: Synthesizing Verifiable Reasoning Data at Scale for Learning Logical Reasoning and Beyond (Liu Jun-teng, 2025) View paper
  - [44] Finllms: A framework for financial reasoning dataset generation with large language models (Ziqiang Yuan, 2024) View paper
  - [47] Figureqa: An annotated figure dataset for visual reasoning (Kahou, 2017) View paper
  - [48] Measuring abstract reasoning in neural networks (David G. T. Barrett, 2018) View paper
  - Data Extraction and Curation (4 papers)
  - [10] Megascience: Pushing the frontiers of post-training datasets for science reasoning (Wang Zeng-zhi, 2025) View paper
  - [18] Openwebmath: An open dataset of high-quality mathematical web text (Paster, 2023) View paper
  - [35] ThoughtSource: A central hub for large language model reasoning data (Simon Ott, 2023) View paper
  - [39] Scaling physical reasoning with the physics dataset (Zheng, 2025) View paper
  - Model Distillation and Trace Generation ★ (5 papers)
  - [0] OpenThoughts: Data Recipes for Reasoning Models (Anon et al., 2026) View paper
  - [6] Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset (Ivan Moshkov, 2025) View paper
  - [13] Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning (Sun Yu, 2025) View paper
  - [14] OpenRTLSet: A Fully Open-Source Dataset for Large Language Model-based Verilog Module Design (Jinghua Wang, 2025) View paper
  - [24] 1.4 million open-source distilled reasoning dataset to empower large language model training (Zhao Han, 2025) View paper
  - Multi-Agent and Iterative Refinement (3 papers)
  - [29] mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans (Sakai Yusuke, 2024) View paper
  - [33] Mathfimer: Enhancing mathematical reasoning by expanding reasoning steps through fill-in-the-middle task (Yan, 2025) View paper
  - [34] OpenMMReasoner: Pushing the Frontiers for Multimodal Reasoning with an Open and General Recipe (Kaichen Zhang, 2025) View paper
- Domain-Specific Reasoning Datasets
  - Mathematical and Coding Reasoning (3 papers)
  - [20] MiMo: Unlocking the Reasoning Potential of Language Model--From Pretraining to Posttraining (- -, 2025) View paper

  - [25] When do program-of-thought works for reasoning? (Bi, 2024) View paper
  - [37] Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models (Albalak, 2025) View paper
  - Scientific and Physical Reasoning (2 papers)
  - [16] Sci-reason: A dataset with chain-of-thought rationales for complex multimodal reasoning in academic areas (MA Chenghao, 2025) View paper
  - [43] PhyX: Does Your Model Have the" Wits" for Physical Reasoning? (Shen Hui, 2025) View paper
  - Medical Reasoning (1 papers)
  - [36] MedCaseReasoning: Evaluating and learning diagnostic reasoning from clinical case reports (Wu Kevin, 2025) View paper
  - Multimodal Reasoning (8 papers)
  - [4] Bigdocs: An open dataset for training multimodal models on document and code tasks (Rodriguez, 2024) View paper
  - [15] Bigdocs: An open and permissively-licensed dataset for training multimodal models on document and code tasks (Rodriguez Juan, 2024) View paper
  - [27] Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding (Lahoud, 2025) View paper
  - [30] Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space (Zhang, 2025) View paper
  - [31] Toolvqa: A dataset for multi-step reasoning vqa with external tools (Yin Shao-feng, 2025) View paper
  - [40] ChartMuseum: Testing Visual Reasoning Capabilities of Large Vision-Language Models (Tang Liyan, 2025) View paper
  - [42] Remi: A dataset for reasoning with multiple images (Kazemi, 2024) View paper
  - [45] PuzzleWorld: A Benchmark for Multimodal, Open-Ended Reasoning in Puzzlehunts (Li HengZhi, 2025) View paper
  - Structured Data and Tool-Augmented Reasoning (4 papers)
  - [8] Tart: An open-source tool-augmented framework for explainable table-based reasoning (Lu Xinyuan, 2025) View paper
  - [17] Openrt: An open-source framework for reasoning over tabular data (Yilun Zhao, 2023) View paper
  - [26] Toolqa: A dataset for llm question answering with external tools (Zhuang, 2023) View paper
  - [50] Finqa: A dataset of numerical reasoning over financial data (Chen Zhiyu, 2021) View paper
- General Reasoning Datasets and Benchmarks
  - Logical and Deductive Reasoning (6 papers)
  - [7] Justlogic: A comprehensive benchmark for evaluating deductive reasoning in large language models (Michael K. Chen, 2025) View paper
  - [12] Logiqa 2.0â□□an improved dataset for logical reasoning in natural language understanding (Hanmeng Liu, 2023) View paper
  - [21] Reclor: A reading comprehension dataset requiring logical reasoning (Weihao Yu, 2020) View paper
  - [23] Logiqa: A challenge dataset for machine reading comprehension with logical reasoning (Liu Jian, 2020) View paper
  - [38] Boardgameqa: A dataset for natural language reasoning with contradictory information (Kazemi, 2023) View paper
  - [41] Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models (Luo Man, 2023) View paper
  - Commonsense and Multi-Hop Reasoning (2 papers)
  - [5] Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps (Xanh Ho, 2020) View paper
  - [46] Naturalreasoning: Reasoning in the wild with 2.8 m challenging questions (Yuan, 2025) View paper
  - Specialized Context Reasoning (1 papers)
  - [22] Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning (Shihao Wang, 2025) View paper
- Training Frameworks and Model Architectures
  - Reinforcement Learning Frameworks (2 papers)
  - [1] Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model (Hu, 2025) View paper
  - [2] Openr: An open source framework for advanced reasoning with large language models (Wang Jun, 2024) View paper
  - Model Architecture Innovations (2 papers)
  - [3] Llama-nemotron: Efficient reasoning models (Levy, 2025) View paper
  - [32] Typhoon T1: An Open Thai Reasoning Model (Taveekitworachai, 2025) View paper
  - Reasoning Process Analysis (3 papers)
  - [19] Unlocking the mysteries of OpenAI o1: A survey of the reasoning abilities of large language models (G Wang, 2025) View paper
  - [28] Long is more important than difficult for training reasoning models (Shen Si, 2025) View paper
  - [49] Rationalyst: Pre-training process-supervision for improving reasoning (Dongwei Jiang, 2025) View paper

Narrative

Core task: creating open-source datasets for training reasoning models. The field organizes around four main branches that reflect different aspects of dataset development. Dataset Construction Methodologies focuses on techniques for generating high-quality reasoning traces, including model distillation approaches that extract intermediate steps from capable models, synthetic data generation methods, and human annotation frameworks. Domain-Specific Reasoning Datasets targets specialized areas such as mathematics, science, medicine, and finance, where domain expertise shapes both problem formulation and solution strategies. General Reasoning Datasets and Benchmarks encompasses broader logical reasoning, multi-hop question answering, and cross-domain evaluation suites that test transferable reasoning skills. Training Frameworks and Model Architectures addresses how datasets integrate with learning paradigms, including reinforcement learning from reasoning traces and architectural innovations that leverage structured thought processes. Representative works like OpenThoughts[0] and Open Reasoner Zero[1] illustrate distillation-based construction, while Llama Nemotron[3] and Openr Framework[2] demonstrate end-to-end training pipelines.

A central tension across these branches involves balancing scale, diversity, and trace quality: distillation methods can rapidly produce large volumes of reasoning steps but may inherit biases from teacher models, whereas human-curated datasets offer higher fidelity at greater cost. Within the Model Distillation and Trace Generation cluster, OpenThoughts[0] emphasizes extracting diverse reasoning patterns from frontier models to create broadly applicable training data, positioning itself alongside efforts like Distilled Reasoning Million[24] that prioritize volume and ReasonMed[13] that targets domain adaptation. Compared to OpenRTLSet[14], which focuses on hardware-specific reasoning, OpenThoughts[0] pursues more general-purpose trace generation. Meanwhile, works like AIMO Winner[6] demonstrate how competition-driven datasets can push mathematical reasoning boundaries, and JustLogic[7] explores formal logic domains. The ongoing challenge remains how to efficiently scale trace generation while maintaining the step-by-step coherence and correctness that enable models to learn robust reasoning strategies rather than superficial pattern matching.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset

Authors: Ivan Moshkov, Darragh Hanley, Toshniwal, Shubham, Ivan Sorokin, et al. (15 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

This paper presents our winning submission to the AI Mathematical Olympiad - Progress Prize 2 (AIMO-2) competition. Our recipe for building state-of-the-art mathematical reasoning models relies on three key pillars. First, we create a large-scale dataset comprising 540K unique high-quality math problems, including olympiad-level problems, and their 3.2M long-reasoning solutions. Second, we develop a novel method to integrate code execution with long reasoning models through iterative training, g...

Relationship Analysis

Both papers belong to the Model Distillation and Trace Generation category, focusing on generating reasoning traces by distilling from stronger models with verification and quality filtering procedures. They overlap in their approach to creating open-source reasoning datasets through teacher model distillation (using DeepSeek-R1 and QwQ-32B), systematic quality filtering, and multiple answer sampling per question. The key differences are that the original paper (OpenThoughts) conducts 1,000+ controlled experiments to systematically investigate each pipeline step across math, code, and science domains, while the candidate paper (AIMO-2 solution) focuses specifically on mathematical reasoning with emphasis on Tool-Integrated Reasoning (TIR) that integrates code execution, and includes a novel Generative Solution Selection (GenSelect) method for choosing the best solution from multiple candidates.

2. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning

Authors: Sun Yu, Qian Xing-yu, Yu Sun, Xu Wei-wen, Xingyu Qian, et al. (24 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Reasoning-based large language models have excelled in mathematics and programming, yet their potential in knowledge-intensive medical question answering remains underexplored and insufficiently validated in clinical contexts. To bridge this gap, we introduce ReasonMed, the largest medical reasoning dataset to date, comprising 370k high-quality examples distilled from 1.75 million initial reasoning paths generated by complementary LLMs and curated through a cost-efficient easy-medium-difficult (...

Relationship Analysis

Both papers belong to the Model Distillation and Trace Generation category, using stronger teacher models to generate reasoning traces for training smaller reasoning models. They overlap in employing multi-model distillation, verification procedures, and quality filtering to create large-scale reasoning datasets. The key difference is that the original paper (OpenThoughts) focuses on systematic pipeline optimization across math, code, and science domains through 1,000+ controlled experiments to identify optimal data generation strategies, while the candidate paper (ReasonMed) specializes in medical reasoning with a multi-agent system and difficulty-based refinement pipeline (easy-medium-difficult) tailored specifically for knowledge-intensive medical QA tasks.

3. OpenRTLSet: A Fully Open-Source Dataset for Large Language Model-based Verilog Module Design

Authors: Jinghua Wang, Lily Jiaxin Wan, Sanjana Pingali, Scott Smith, Manvi Jha, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

OpenRTLSet1 introduces the largest fully open-source dataset for hardware design, offering over 127,000 diverse Verilog code samples to the research community and industry. Our dataset uniquely combines Verilog code from GitHub repositories (98k modules), VHDL translations (5k modules), and synthesizable C/C++ translations (24k modules), all freely accessible without proprietary restrictions. Using the reasoning model DeepSeek-R1, we generated paired natural language descriptions for each code s...

Relationship Analysis

Both papers belong to the Model Distillation and Trace Generation category, using stronger reasoning models to generate training data through distillation. They overlap in their approach of using DeepSeek-R1 as a teacher model to generate reasoning traces for open-source datasets. However, OpenThoughts focuses on creating general reasoning datasets across math, code, and science domains with extensive ablation studies on data curation strategies, while OpenRTLSet specializes exclusively in hardware design (Verilog code generation) by combining multiple code sources and exploring quantization techniques for domain-specific applications.

4. 1.4 million open-source distilled reasoning dataset to empower large language model training

Authors: Zhao Han, Wang, Haotian, Han Zhao, Peng Yi-ping, et al. (17 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The AM-DeepSeek-R1-Distilled is a large-scale dataset with thinking traces for general reasoning tasks, composed of high-quality and challenging reasoning problems. These problems are collected from a multitude of open-source datasets, subjected to semantic deduplication and meticulous cleaning to eliminate test set contamination. All responses within the dataset are distilled from reasoning models (predominantly DeepSeek-R1) and have undergone rigorous verification procedures. Mathematical prob...

Relationship Analysis

Both papers belong to the Model Distillation and Trace Generation category, focusing on creating reasoning datasets by distilling from stronger teacher models with verification procedures. They overlap in their approach of using DeepSeek-R1 as a primary teacher model, applying verification methods (answer checking, test cases), and targeting math, code, and science domains. The key differences are that the original paper (OpenThoughts) conducts 1,000+ systematic ablation experiments to optimize each pipeline step (question sourcing, mixing, filtering, deduplication, answer sampling), discovering that QwQ-32B outperforms DeepSeek-R1 as a teacher and that mixing fewer high-quality sources works better, while the candidate paper (AM) focuses on scaling to 1.4M examples with emphasis on semantic deduplication, difficulty-based filtering, and reward model verification without the extensive ablation study.

Contributions Analysis

Overall novelty summary. The paper contributes OpenThoughts3-1.2M, a systematically refined dataset for training reasoning models, and OpenThinker3-7B, which achieves state-of-the-art results on AIME, LiveCodeBench, and GPQA Diamond. It resides in the 'Model Distillation and Trace Generation' leaf, which contains five papers total, including the original work. This leaf sits within the broader 'Dataset Construction Methodologies' branch, indicating a moderately populated research direction focused on extracting reasoning traces from stronger models rather than synthetic generation or human annotation.

The taxonomy reveals neighboring leaves such as 'Synthetic Data Generation' (five papers) and 'Multi-Agent and Iterative Refinement' (three papers), both under the same parent branch. The paper's distillation-based approach contrasts with template-driven synthesis methods in the former and multi-agent verification strategies in the latter. Its sibling papers include OpenThoughts (earlier version), Distilled Reasoning Million, ReasonMed, and OpenRTLSet, which collectively explore distillation at scale, domain adaptation, and hardware-specific reasoning. The taxonomy structure suggests this is an active but not overcrowded subfield within dataset construction.

Among 30 candidates examined, none clearly refute the three main contributions. For the OpenThoughts3-1.2M dataset and pipeline, 10 candidates were reviewed with zero refutable overlaps. Similarly, the OpenThinker3-7B model and empirical insights on data curation each examined 10 candidates without finding prior work that directly overlaps. This limited search scope suggests the specific combination of systematic pipeline investigation (1,000+ experiments), QwQ-32B distillation, and the resulting performance gains may be novel within the examined literature, though the analysis does not cover the full field.

Based on top-30 semantic matches, the work appears to advance distillation-based dataset construction through systematic experimentation and achieves notable empirical results. However, the search scope is constrained, and the taxonomy shows this is an established research direction with multiple concurrent efforts. The novelty likely lies in the methodological rigor of pipeline optimization and the specific performance improvements demonstrated, rather than introducing entirely new conceptual approaches to reasoning dataset creation.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: OpenThoughts3-1.2M dataset and systematic data generation pipeline

Description: The authors develop a systematic data generation pipeline through over 1,000 controlled experiments, investigating each step including question sourcing, mixing, filtering, deduplication, answer sampling, answer filtering, and teacher model selection. This pipeline produces OpenThoughts3-1.2M, a dataset of 1.2 million examples for training reasoning models across math, code, and science domains.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. SwS: Self-aware Weakness-driven Problem Synthesis in Reinforcement Learning for LLM Reasoning

URL: View paper

Brief Assessment

SwS[66] focuses on reinforcement learning with problem synthesis driven by model weaknesses, not on systematic data generation pipelines for creating large-scale reasoning datasets through controlled experiments on question sourcing, filtering, and teacher model selection.

2. Unicorn: Text-Only Data Synthesis for Vision Language Model Training

URL: View paper

Brief Assessment

Unicorn[63] focuses on synthesizing multimodal training data from text for vision-language models, not on creating reasoning traces for mathematical, coding, and scientific reasoning tasks. The domains, methodologies, and objectives are fundamentally different.

3. Spatialrgpt: Grounded spatial reasoning in vision-language models

URL: View paper

Brief Assessment

SpatialRGPT[62] focuses on spatial reasoning in vision-language models using 3D scene graphs and depth information, not on systematic data generation pipelines for training reasoning models across math, code, and science domains.

4. TARGA: Targeted Synthetic Data Generation for Practical Reasoning over Structured Data

URL: View paper

Brief Assessment

TARGA[65] focuses on targeted synthetic data generation for semantic parsing over structured data (knowledge bases, databases), not on training reasoning models for math, code, and science domains. The candidate addresses a different problem domain with different methodology.

5. Long is more important than difficult for training reasoning models

URL: View paper

Brief Assessment

Long Training[28] focuses on a different research question—whether reasoning length or problem difficulty matters more for training. While both papers involve data generation for reasoning models, Long Training[28] does not demonstrate prior work on the systematic pipeline described in the original paper (question sourcing, mixing, filtering, deduplication, answer sampling, answer filtering, and teacher model selection through 1,000+ controlled experiments). The candidate's synthesis method (concatenating two problems) differs fundamentally from the original's multi-stage pipeline approach.

6. Synthetic data generation & multi-step rl for reasoning & tool use

URL: View paper

Brief Assessment

Synthetic RL[61] focuses on multi-step reinforcement learning for reasoning and tool use, not on systematic data generation pipelines for training reasoning models. The candidate's methodology centers on step-wise RL optimization and synthetic trajectory generation for tool use scenarios, which is technically distinct from the original paper's systematic investigation of data generation pipeline components (question sourcing, mixing, filtering, deduplication, answer sampling, teacher model selection) through 1,000+ controlled experiments.

7. SynLogic: Synthesizing Verifiable Reasoning Data at Scale for Learning Logical Reasoning and Beyond

URL: View paper

Brief Assessment

SynLogic[11] focuses on logical reasoning tasks (35 diverse logic tasks like Sudoku, Game of 24) rather than general reasoning across math, code, and science domains. The pipeline designs and verification methods are domain-specific to logic problems, not the multi-domain systematic investigation described in the original contribution.

8. Towards large reasoning models: A survey of reinforced reasoning with large language models

URL: View paper

Brief Assessment

Large Reasoning Models[51] is a survey paper that reviews existing methods for training reasoning models but does not present a novel data generation pipeline or dataset. It discusses general approaches like MCTS and automated data construction but does not claim priority over specific implementations.

9. Slr: Automated synthesis for scalable logical reasoning

URL: View paper

Brief Assessment

SLR[64] focuses on automated synthesis for logical reasoning tasks using symbolic validation programs, not on systematic data generation pipelines for training reasoning models across math, code, and science domains as described in the original paper.

10. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps

URL: View paper

Brief Assessment

Multi-hop QA[5] focuses on constructing a multi-hop question answering dataset with evidence paths for reasoning evaluation, not on systematic data generation pipelines for training reasoning models across math, code, and science domains.

Contribution 2: OpenThinker3-7B state-of-the-art reasoning model

Description: The authors train OpenThinker3-7B by fine-tuning Qwen2.5-7B-Instruct on their OpenThoughts3-1.2M dataset, achieving state-of-the-art performance among open-data reasoning models at the 7B scale, with improvements of 15.3, 17.2, and 20.5 percentage points over DeepSeek-R1-Distill-Qwen-7B on key benchmarks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Llm reasoning engine: Specialized training for enhanced mathematical reasoning

URL: View paper

Brief Assessment

LLM Reasoning Engine[71] focuses on mathematical reasoning through question paraphrasing and specialized training objectives (rationale re-ranking, mistake identification), not on general reasoning model development or the specific training recipe used for OpenThinker3-7B. The candidate does not demonstrate prior work on the comprehensive data pipeline or the specific achievements claimed.

2. Scitune: Aligning large language models with scientific multimodal instructions

URL: View paper

Brief Assessment

SciTune[76] focuses on aligning language models with scientific multimodal instructions for visual and language understanding tasks, not on general mathematical, coding, and scientific reasoning through long chain-of-thought generation.

3. Solving quantitative reasoning problems with language models

URL: View paper

Brief Assessment

Quantitative Reasoning[72] focuses on mathematical problem-solving using natural language and symbolic manipulation without external tools, whereas the original paper trains reasoning models on diverse datasets (math, code, science) using supervised fine-tuning with long chain-of-thought traces.

4. Advancing reasoning in large language models: Promising methods and approaches

URL: View paper

Brief Assessment

Advancing Reasoning[69] is a survey paper that reviews general methods for enhancing reasoning in LLMs (prompting strategies, architectural innovations, learning paradigms). It does not present a specific 7B-scale reasoning model or training methodology that would refute the novelty of OpenThinker3-7B's fine-tuning approach and benchmark achievements.

5. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct

URL: View paper

Brief Assessment

WizardMath[74] focuses on mathematical reasoning through reinforced evol-instruct without external tools, while the original paper addresses broader reasoning across math, code, and science with a systematic data pipeline approach. The training methodologies and data generation strategies differ fundamentally.

6. Improving large language model fine-tuning for solving math problems

URL: View paper

Brief Assessment

Math Fine-tuning[75] focuses on fine-tuning methods for mathematical problem solving using the MATH dataset, not on creating open-data reasoning models or the specific OpenThoughts3 dataset pipeline that produces OpenThinker3-7B.

7. Program synthesis with large language models

URL: View paper

Brief Assessment

Program Synthesis[73] focuses on synthesizing Python programs from natural language descriptions using large language models, not on training reasoning models for mathematical, coding, and scientific tasks through supervised fine-tuning on reasoning traces.

8. Can large language models detect errors in long chain-of-thought reasoning?

URL: View paper

Brief Assessment

Error Detection CoT[70] focuses on evaluating error detection capabilities in long chain-of-thought reasoning rather than training state-of-the-art reasoning models through fine-tuning on curated datasets.

9. Internlm-math: Open math large language models toward verifiable reasoning

URL: View paper

Brief Assessment

InternLM-Math[67] focuses on mathematical reasoning with formal verification using Lean, while the original paper targets general reasoning across math, code, and science domains using supervised fine-tuning on distilled reasoning traces. The approaches and evaluation settings differ substantially.

10. How abilities in large language models are affected by supervised fine-tuning data composition

URL: View paper

Brief Assessment

SFT Data Composition[68] focuses on how data composition ratios affect multiple abilities (math, code, general alignment) during supervised fine-tuning, not on training a specific state-of-the-art reasoning model or achieving benchmark improvements through dataset curation.

Contribution 3: Empirical insights on reasoning data curation strategies

Description: Through systematic ablation studies, the authors discover several key findings: sampling multiple answers per question (16×) effectively increases dataset scale; weaker teacher models like QwQ-32B can outperform stronger ones like DeepSeek-R1; answer filtering provides minimal benefit; selecting questions from 1-2 high-quality sources outperforms mixing many sources; and LLM-based question filtering (difficulty, response length) outperforms classical methods like fastText.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Syndarin: Synthesising datasets for automated reasoning in low-resource languages

URL: View paper

Brief Assessment

Syndarin[55] focuses on creating QA datasets for low-resource languages through parallel content mining and translation validation, not on data curation strategies for improving reasoning capabilities in language models through techniques like answer sampling, teacher model selection, or question filtering.

2. PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models

URL: View paper

Brief Assessment

PHYBench[58] focuses on benchmark design and evaluation methodology for physics reasoning tasks, not on data curation strategies for training reasoning models. The candidate addresses evaluation challenges rather than training data generation pipelines.

3. Learning Theorem Rationale for Improving the Mathematical Reasoning Capability of Large Language Models

URL: View paper

Brief Assessment

Theorem Rationale[56] focuses on teaching LLMs to explicitly select and apply mathematical theorems to problems, not on general data curation strategies like sampling multiple answers, teacher model selection, or question filtering methods that the original paper investigates.

4. Improve vision language model chain-of-thought reasoning

URL: View paper

Brief Assessment

Vision CoT[60] focuses on vision-language model reasoning through chain-of-thought augmentation and reinforcement learning for visual question answering tasks, not on general language model reasoning data curation strategies across math, code, and science domains as explored in the original paper.

5. Specializing smaller language models towards multi-step reasoning

URL: View paper

Brief Assessment

Smaller Model Reasoning[59] focuses on distilling reasoning abilities from large models (GPT-3.5) to smaller T5 models (≤11B) for math reasoning tasks, not on systematic data curation strategies across multiple domains (math, code, science) as in the original paper. The candidate explores model specialization through distillation but does not conduct the extensive ablation studies (1,000+ experiments) on data sourcing, filtering, mixing, and teacher model selection that form the core of the original contribution.

6. Large language models can self-improve

URL: View paper

Brief Assessment

Self-Improve[54] focuses on self-training methods where models generate and filter their own training data using chain-of-thought prompting and self-consistency. The paper does not systematically investigate data curation strategies like sampling multiple answers per question, teacher model selection, question filtering methods, or question source mixing—which are the core empirical insights claimed in the original paper.

7. Towards large reasoning models: A survey of reinforced reasoning with large language models

URL: View paper

Brief Assessment

The candidate paper surveys existing data curation techniques (e.g., MCTS simulation, teacher model selection) but does not present original empirical findings that would refute the novelty of the specific ablation studies and insights reported in the original paper.

8. Large language models are reasoning teachers

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

9. Lisa: Reasoning segmentation via large language model

URL: View paper

Brief Assessment

LISA[52] focuses on reasoning segmentation for computer vision tasks, not on data curation strategies for improving reasoning capabilities in language models. The paper addresses a completely different domain (vision-language models for segmentation) rather than general reasoning model training.

10. Towards reasoning ability of small language models

URL: View paper

Brief Assessment

Small Language Reasoning[57] focuses on evaluating reasoning capabilities of small language models through benchmarking and compression techniques, not on data curation strategies for training reasoning models.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] OpenThoughts: Data Recipes for Reasoning Models View paper
- [1] Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model View paper
- [2] Openr: An open source framework for advanced reasoning with large language models View paper
- [3] Llama-nemotron: Efficient reasoning models View paper
- [4] Bigdocs: An open dataset for training multimodal models on document and code tasks View paper
- [5] Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps View paper
- [6] Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset View paper
- [7] Justlogic: A comprehensive benchmark for evaluating deductive reasoning in large language models View paper
- [8] Tart: An open-source tool-augmented framework for explainable table-based reasoning View paper
- [9] Key-point-driven data synthesis with its enhancement on mathematical reasoning View paper
- [10] Megascience: Pushing the frontiers of post-training datasets for science reasoning View paper
- [11] SynLogic: Synthesizing Verifiable Reasoning Data at Scale for Learning Logical Reasoning and Beyond View paper
- [12] Logiqa 2.0â□□an improved dataset for logical reasoning in natural language understanding View paper
- [13] Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning View paper
- [14] OpenRTLSet: A Fully Open-Source Dataset for Large Language Model-based Verilog Module Design View paper
- [15] Bigdocs: An open and permissively-licensed dataset for training multimodal models on document and code tasks View paper
- [16] Sci-reason: A dataset with chain-of-thought rationales for complex multimodal reasoning in academic areas View paper
- [17] Openrt: An open-source framework for reasoning over tabular data View paper
- [18] Openwebmath: An open dataset of high-quality mathematical web text View paper
- [19] Unlocking the mysteries of OpenAI o1: A survey of the reasoning abilities of large language models View paper
- [20] MiMo: Unlocking the Reasoning Potential of Language Model--From Pretraining to Posttraining View paper
- [21] Reclor: A reading comprehension dataset requiring logical reasoning View paper
- [22] Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning View paper
- [23] Logiqa: A challenge dataset for machine reading comprehension with logical reasoning View paper
- [24] 1.4 million open-source distilled reasoning dataset to empower large language model training View paper
- [25] When do program-of-thought works for reasoning? View paper
- [26] Toolqa: A dataset for llm question answering with external tools View paper
- [27] Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding View paper
- [28] Long is more important than difficult for training reasoning models View paper
- [29] mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans View paper
- [30] Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space View paper
- [31] Toolvqa: A dataset for multi-step reasoning vqa with external tools View paper
- [32] Typhoon T1: An Open Thai Reasoning Model View paper
- [33] Mathfimer: Enhancing mathematical reasoning by expanding reasoning steps through fill-in-the-middle task View paper
- [34] OpenMMReasoner: Pushing the Frontiers for Multimodal Reasoning with an Open and General Recipe View paper
- [35] ThoughtSource: A central hub for large language model reasoning data View paper
- [36] MedCaseReasoning: Evaluating and learning diagnostic reasoning from clinical case reports View paper
- [37] Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models View paper
- [38] Boardgameqa: A dataset for natural language reasoning with contradictory information View paper
- [39] Scaling physical reasoning with the physics dataset View paper
- [40] ChartMuseum: Testing Visual Reasoning Capabilities of Large Vision-Language Models View paper
- [41] Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models View paper
- [42] Remi: A dataset for reasoning with multiple images View paper
- [43] PhyX: Does Your Model Have the" Wits" for Physical Reasoning? View paper
- [44] Finllms: A framework for financial reasoning dataset generation with large language models View paper
- [45] PuzzleWorld: A Benchmark for Multimodal, Open-Ended Reasoning in Puzzlehunts View paper
- [46] Naturalreasoning: Reasoning in the wild with 2.8 m challenging questions View paper
- [47] Figureqa: An annotated figure dataset for visual reasoning View paper
- [48] Measuring abstract reasoning in neural networks View paper
- [49] Rationalyst: Pre-training process-supervision for improving reasoning View paper
- [50] Finqa: A dataset of numerical reasoning over financial data View paper
- [51] Towards large reasoning models: A survey of reinforced reasoning with large language models View paper
- [52] Lisa: Reasoning segmentation via large language model View paper
- [53] Large language models are reasoning teachers View paper
- [54] Large language models can self-improve View paper

- [55] Syndarin: Synthesising datasets for automated reasoning in low-resource languages View paper
- [56] Learning Theorem Rationale for Improving the Mathematical Reasoning Capability of Large Language Models View paper
- [57] Towards reasoning ability of small language models View paper
- [58] PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models View paper
- [59] Specializing smaller language models towards multi-step reasoning View paper
- [60] Improve vision language model chain-of-thought reasoning View paper
- [61] Synthetic data generation & multi-step rl for reasoning & tool use View paper
- [62] Spatialrgpt: Grounded spatial reasoning in vision-language models View paper
- [63] Unicorn: Text-Only Data Synthesis for Vision Language Model Training View paper
- [64] Slr: Automated synthesis for scalable logical reasoning View paper
- [65] TARGA: Targeted Synthetic Data Generation for Practical Reasoning over Structured Data View paper
- [66] SwS: Self-aware Weakness-driven Problem Synthesis in Reinforcement Learning for LLM Reasoning View paper
- [67] Internlm-math: Open math large language models toward verifiable reasoning View paper
- [68] How abilities in large language models are affected by supervised fine-tuning data composition View paper
- [69] Advancing reasoning in large language models: Promising methods and approaches View paper
- [70] Can large language models detect errors in long chain-of-thought reasoning? View paper
- [71] Llm reasoning engine: Specialized training for enhanced mathematical reasoning View paper
- [72] Solving quantitative reasoning problems with language models View paper
- [73] Program synthesis with large language models View paper
- [74] Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct View paper
- [75] Improving large language model fine-tuning for solving math problems View paper
- [76] Scitune: Aligning large language models with scientific multimodal instructions View paper