Novelty Assessment Report

Paper: AceReason-Nemotron 1.1: Advancing Math and Code Reasoning through SFT and RL Synergy
PDF URL: https://openreview.net/pdf?id=IaEqjWXd1d
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

In this work, we investigate the synergy between supervised fine-tuning (SFT) and reinforcement learning (RL) in developing strong reasoning models. We begin by curating the SFT training data through two scaling strategies: increasing the number of collected prompts and the number of generated responses per prompt. Both approaches yield notable improvements in reasoning performance, with scaling the number of prompts resulting in more substantial gains. We then explore the following questions regarding the synergy between SFT and RL: (i) Does a stronger SFT model consistently lead to better final performance after large-scale RL training? (ii) How can we determine an appropriate sampling temperature during RL training to effectively balance exploration and exploitation for a given SFT initialization?
Our findings suggest that (i) holds true, provided effective RL training is conducted, particularly when the sampling temperature is carefully chosen to maintain the temperature-adjusted entropy around 0.3, a setting that strikes a good balance between exploration and exploitation. Notably, the performance gap between initial SFT models narrows significantly throughout the RL process. Built on a strong SFT foundation and SFT–RL synergy, our AceReason-Nemotron-1.1 7B model significantly outperforms AceReason-Nemotron-1.0 and achieves new state-of-the-art performance among Qwen2.5-7B-based reasoning models on challenging math and code benchmarks, thereby demonstrating the effectiveness of our post-training recipe.

Core Task Landscape

This paper addresses: **Synergy between Supervised Fine-Tuning and Reinforcement Learning for Reasoning Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

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

- **Integration Frameworks and Training Paradigms**
- **Theoretical Foundations and Comparative Analysis**
- **Application Domains and Task-Specific Methods**
- **Reinforcement Learning Techniques and Optimization**
- **Survey and Review Studies**
- **Fine-Tuning Impact and Side Effects**

Complete Taxonomy Tree

- Synergy between Supervised Fine-Tuning and Reinforcement Learning for Reasoning Models Survey Taxonomy
- Integration Frameworks and Training Paradigms
  - Unified and Dynamic Integration Methods (5 papers)
  - [2] On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting (Zhang Wenhao, 2025) View paper
  - [3] Step-wise Adaptive Integration of Supervised Fine-tuning and Reinforcement Learning for Task-Specific LLMs (Chen Jack, 2025) View paper
  - [7] SRFT: A Single-Stage Method with Supervised and Reinforcement Fine-Tuning for Reasoning (Fu, 2025) View paper
  - [11] UFT: Unifying Supervised and Reinforcement Fine-Tuning (Liu Ming-Yang, 2025) View paper
  - [37] Beyond Two-Stage Training: Cooperative SFT and RL for LLM Reasoning (Chen Liang, 2025) View paper
  - Sequential Training Strategies (4 papers)
  - [15] Learning What Reinforcement Learning Can't: Interleaved Online Fine-Tuning for Hardest Questions (Ma Lu, 2025) View paper
  - [20] SuperRL: Reinforcement Learning with Supervision to Boost Language Model Reasoning (Liu Yihao, 2025) View paper
  - [26] Beyond Two-Stage Training: Integrating SFT and RL for Improved Reasoning in LLMs (C Liang, 2025) View paper
  - [42] ARES: Alternating Reinforcement Learning and Supervised Fine-Tuning for Enhanced Multi-Modal Chain-of-Thought Reasoning Through Diverse AI Feedback (Byun, 2024) View paper
  - Prefix-Based Hybrid Approaches (2 papers)
  - [14] Reasoning with reinforced functional token tuning (Yao Qi, 2025) View paper
  - [25] Blending Supervised and Reinforcement Fine-Tuning with Prefix Sampling (Huang, 2025) View paper
- Theoretical Foundations and Comparative Analysis
  - Mechanistic Analysis and Learning Dynamics (3 papers)
  - [6] Demystifying long chain-of-thought reasoning in llms (Tong, 2025) View paper
  - [43] How Much Backtracking is Enough? Exploring the Interplay of SFT and RL in Enhancing LLM Reasoning (Wang Junlin, 2025) View paper
  - [46] RL Squeezes, SFT Expands: A Comparative Study of Reasoning LLMs (Kohsei Matsutani, 2025) View paper
  - Comparative Performance Studies (4 papers)

- [9] Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering (Li Gang, 2025) View paper
- [29] Reassessing the Role of Supervised Fine-Tuning: An Empirical Study in VLM Reasoning (Yongcan Yu, 2025) View paper
- [33] The Synergy Dilemma of Long-CoT SFT and RL: Investigating Post-Training Techniques for Reasoning VLMs (Chen, 2025) View paper
- [47] SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models (Tu, 2025) View paper
- Forgetting and Recovery Mechanisms (2 papers)
- [40] RL Fine-Tuning Heals OOD Forgetting in SFT (Luan, 2025) View paper
- [41] BREAD: Branched Rollouts from Expert Anchors Bridge SFT & RL for Reasoning (Zhang Xuechen, 2025) View paper

- Application Domains and Task-Specific Methods
  - Vision-Language Reasoning
  - Visual Reinforcement Fine-Tuning Frameworks (4 papers)
    - [1] Visual-rft: Visual reinforcement fine-tuning (Liu Zi-yu, 2025) View paper
    - [4] Reason-rft: Reinforcement fine-tuning for visual reasoning (Tan Huajie, 2025) View paper
    - [8] Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models (Tan Huajie, 2025) View paper
    - [27] Grounded Reinforcement Learning for Visual Reasoning (Sarch, 2025) View paper
  - General Multimodal RL Methods (5 papers)
    - [12] Fine-tuning large vision-language models as decision-making agents via reinforcement learning (Zhai, 2024) View paper
    - [23] Advancing Multimodal Reasoning via Reinforcement Learning with Cold Start (Wei Lai, 2025) View paper
    - [31] ViSurf: Visual Supervised-and-Reinforcement Fine-Tuning for Large Vision-and-Language Models (Liu Yuqi, 2025) View paper
    - [38] Omni-AutoThink: Adaptive Multimodal Reasoning via Reinforcement Learning (Dongchao Yang, 2025) View paper
    - [49] Metis-RISE: RL Incentivizes and SFT Enhances Multimodal Reasoning Model Learning (Qiu Haibo, 2025) View paper
  - Long Video Reasoning (1 papers)
    - [48] Scaling RL to Long Videos (Chen, 2025) View paper
  - Domain-Specific Visual Applications (2 papers)
    - [32] UAV-VL-R1: Generalizing Vision-Language Models via Supervised Fine-Tuning and Multi-Stage GRPO for UAV Visual Reasoning (Mei Haibo, 2025) View paper
    - [35] MIRG-RL: Multi-Image Reasoning and Grounding with Reinforcement Learning (Zheng Li-hao, 2025) View paper
  - Mathematical and Code Reasoning ★ (6 papers)
  - [0] AceReason-Nemotron 1.1: Advancing Math and Code Reasoning through SFT and RL Synergy (Anon et al., 2026) View paper
  - [5] Teaching large language models to reason with reinforcement learning (Havrilla, 2024) View paper
  - [10] Bridging Supervised Learning and Reinforcement Learning in Math Reasoning (Chen Huayu, 2025) View paper
  - [13] Reft: Reasoning with reinforced fine-tuning (Luong, 2024) View paper
  - [17] Unlock the Correlation between Supervised Fine-Tuning and Reinforcement Learning in Training Code Large Language Models (Chen Jie, 2024) View paper
  - [28] G1: Teaching LLMs to Reason on Graphs with Reinforcement Learning (Guo Xiaojun, 2025) View paper
  - Specialized Domain Applications (3 papers)
  - [16] ERank: Fusing Supervised Fine-Tuning and Reinforcement Learning for Effective and Efficient Text Reranking (Cai, 2025) View paper
  - [21] Enhancing LLMs' Reasoning-Intensive Multimedia Search Capabilities through Fine-Tuning and Reinforcement Learning (Li, 2025) View paper
  - [24] Reasoning-Table: Exploring Reinforcement Learning for Table Reasoning (Lei Fang-yu, 2025) View paper
- Reinforcement Learning Techniques and Optimization
  - Policy Optimization and Reward Modeling (5 papers)
  - [22] Process-Supervised Reinforcement Learning for Interactive Multimodal Tool-Use Agents (Tan, 2025) View paper
  - [30] Anchored Supervised Fine-Tuning (Zhu He, 2025) View paper
  - [36] Effective Reinforcement Learning for Reasoning in Language Models (Li Shuo, 2025) View paper
  - [44] Offline RL by Reward-Weighted Fine-Tuning for Conversation Optimization (Mukherjee, 2025) View paper
  - [50] Reinforcement Learning Fine-Tuning of Language Model for Instruction Following and Math Reasoning (Yifu Han, 2025) View paper
  - Token-Efficient RL Methods (1 papers)
  - [39] Token-Efficient RL for LLM Reasoning (Alan, 2025) View paper
  - Exploration and Entropy-Based Methods (1 papers)
  - [34] Reasoning through Exploration: A Reinforcement Learning Framework for Robust Function Calling (Hao Bing-guang, 2025) View paper
- Survey and Review Studies (2 papers)
  - [18] Reasoning beyond limits: Advances and open problems for llms (Ferrag, 2025) View paper
  - [19] Advancing reasoning in large language models: Promising methods and approaches (Patil Avinash, 2025) View paper
- Fine-Tuning Impact and Side Effects (1 papers)
  - [45] On the Impact of Fine-Tuning on Chain-of-Thought Reasoning (Agarwal, 2024) View paper

Narrative

Core task: Synergy between supervised fine-tuning and reinforcement learning for reasoning models. The field explores how supervised fine-tuning (SFT) and reinforcement learning (RL) can be combined to enhance reasoning capabilities in language and multimodal models. The taxonomy reveals several major branches: Integration Frameworks and Training Paradigms examine how to orchestrate SFT and RL stages, including sequential pipelines and interleaved approaches like Interleaved Online Fine-Tuning[15]; Theoretical Foundations and Comparative Analysis investigate when and why each method excels, as seen in works like On-policy Off-policy Harmony[2] and Synergy Dilemma CoT[33]; Application Domains and Task-Specific Methods focus on deploying these techniques in mathematical reasoning (Bridging SFT RL Math[10]), code generation (SFT RL Correlation Code[17]), and vision-language tasks (Visual RFT[1], Reason RFT VLM[8]); Reinforcement Learning Techniques and Optimization develop novel RL algorithms and reward mechanisms such as SRFT[7] and Token-Efficient RL[39]; while Fine-Tuning Impact and Side Effects study how SFT influences downstream RL performance, exemplified by RL Squeezes SFT Expands[46] and Impact Fine-Tuning CoT[45].

A particularly active line of work centers on mathematical and code reasoning, where researchers debate optimal training sequences and the relative contributions of SFT versus RL. Teaching LLMs Reason[5] and Bridging SFT RL Math[10] explore foundational strategies for

combining both paradigms in structured reasoning tasks, while works like REFT[13] and Reason RFT[4] propose refined rejection sampling and filtering techniques to improve data quality before RL training. AceReason Nemotron[0] situates itself within this mathematical and code reasoning cluster, emphasizing adaptive integration strategies that balance SFT's ability to provide strong initial reasoning patterns with RL's capacity for exploration and reward-driven refinement. Compared to Step-wise Adaptive Integration[3], which dynamically adjusts training phases, and SFT RL Correlation Code[17], which analyzes correlations in code tasks, AceReason Nemotron[0] focuses on achieving synergy through careful orchestration of both methods to maximize reasoning performance across diverse problem types.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. Teaching large language models to reason with reinforcement learning

Authors: Havrilla, Alex, Du Yuqing, Alex Havrilla, Raparthy, et al. (24 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Reinforcement Learning from Human Feedback (\textbf{RLHF}) has emerged as a dominant approach for aligning LLM outputs with human preferences. Inspired by the success of RLHF, we study the performance of multiple algorithms that learn from feedback (Expert Iteration, Proximal Policy Optimization (\textbf{PPO}), Return-Conditioned RL) on improving LLM reasoning capabilities. We investigate both sparse and dense rewards provided to the LLM both heuristically and via a learned reward model. We addi...

Relationship Analysis

Both papers belong to the Mathematical and Code Reasoning category, exploring SFT-RL synergy for reasoning tasks with verifiable rewards. They overlap in investigating how supervised fine-tuning and reinforcement learning interact to improve mathematical reasoning capabilities, both using similar benchmarks (GSM8K, MATH) and RL algorithms (PPO, Expert Iteration). The key difference is that the original paper (AceReason-Nemotron 1.1) focuses on systematic scaling of SFT data, temperature-adjusted entropy for exploration-exploitation balance, and stage-wise RL training on math-only and code-only prompts, while the candidate paper provides a broader comparative study of multiple RL algorithms (PPO, Expert Iteration, RCRL) across different reward schemes and model initializations, with emphasis on sample complexity analysis and exploration limitations.

---

2. Bridging Supervised Learning and Reinforcement Learning in Math Reasoning

Authors: Chen Huayu, Zheng, Kaiwen, Huayu Chen, Zhang, et al. (26 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Reinforcement Learning (RL) has played a central role in the recent surge of LLMs' math abilities by enabling self-improvement through binary verifier signals. In contrast, Supervised Learning (SL) is rarely considered for such verification-driven training, largely due to its heavy reliance on reference answers and inability to reflect on mistakes. In this work, we challenge the prevailing notion that self-improvement is exclusive to RL and propose Negative-aware Fine-Tuning (NFT) -- a supervise...

Relationship Analysis

Both papers belong to the Mathematical and Code Reasoning category, applying SFT-RL synergy to enhance reasoning capabilities with verifiable rewards. They overlap in exploring the interplay between supervised fine-tuning and reinforcement learning for math reasoning tasks, using similar base models (Qwen2.5-Math-7B) and evaluation benchmarks (AIME, MATH500). The key difference is that the original paper (AceReason-Nemotron 1.1) focuses on systematic scaling of SFT data, stage-wise RL curriculum with response length extension, and temperature tuning for exploration-exploitation balance, while the candidate paper (Bridging SL and RL) introduces Negative-aware Fine-Tuning (NFT), a supervised learning method that leverages negative feedback through implicit negative policy modeling, and establishes theoretical equivalence between NFT and GRPO in on-policy settings.

---

3. Reft: Reasoning with reinforced fine-tuning

Authors: Luong, Trung Quoc, Zhang Xin-bo, Trung Quoc Luong, Jie, et al. (15 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

One way to enhance the reasoning capability of Large Language Models (LLMs) is to conduct Supervised Fine-Tuning (SFT) using Chain-of-Thought (CoT) annotations. This approach does not show sufficiently strong generalization ability, however, because the training only relies on the given CoT data. In math problem-solving, for example, there is usually only one annotated reasoning path for each question in the training data. Intuitively, it would be better for the algorithm to learn from multiple ...

Relationship Analysis

Both papers belong to the Mathematical and Code Reasoning category, exploring SFT-RL synergy for reasoning tasks with verifiable rewards. They overlap in investigating how supervised fine-tuning and reinforcement learning can be combined to improve mathematical problem-solving capabilities, with both using PPO/GRPO-style RL algorithms after SFT warmup. The key differences are that the original paper (AceReason-Nemotron 1.1) focuses on systematic analysis of SFT-RL synergy through scaling studies, temperature tuning, and stage-wise RL training on 7B models, while the candidate paper (ReFT) presents a simpler two-stage approach (SFT warmup followed by PPO) without the detailed ablation studies or multi-stage curriculum found in the original work.

---

4. Unlock the Correlation between Supervised Fine-Tuning and Reinforcement Learning in Training Code Large Language Models

Authors: Chen Jie, Jie Chen, Han, Xintian, Xintian Han, et al. (12 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Automatic code generation has been a longstanding research topic. With the advancement of general-purpose large language models (LLMs), the ability to code stands out as one important measure to the model's reasoning performance. Usually, a two-stage training paradigm is implemented to obtain a Code LLM, namely the pretraining and the fine-tuning. Within the fine-tuning, supervised fine-tuning (SFT), and reinforcement learning (RL) are often used to improve the model's zero-shot ability. A large...

Relationship Analysis

Both papers belong to the Mathematical and Code Reasoning category, investigating the synergy between SFT and RL for improving reasoning capabilities in verifiable domains. They overlap in exploring how SFT initialization affects RL performance and examining the interplay between these two training stages for code and mathematical reasoning tasks. However, the original paper (AceReason-Nemotron 1.1) focuses on scaling SFT data through prompt collection and response generation, temperature-adjusted entropy for RL exploration-exploitation balance, and stage-wise RL with progressive response length extension (8K to 32K), while the candidate paper emphasizes synthetic data generation from atomic functions, investigates whether RL can be trained from scratch without SFT, and examines overfitting issues when using identical prompts across SFT and RL phases.

---

5. G1: Teaching LLMs to Reason on Graphs with Reinforcement Learning

Authors: Guo Xiaojun, Li Ang, Xiaojun Guo, Wang Yi-fei, Ang Li, et al. (12 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Although Large Language Models (LLMs) have demonstrated remarkable progress, their proficiency in graph-related tasks remains notably limited, hindering the development of truly general-purpose models. Previous attempts, including pretraining graph foundation models or employing supervised fine-tuning, often face challenges such as the scarcity of large-scale, universally represented graph data. We introduce G1, a simple yet effective approach demonstrating that Reinforcement Learning (RL) on sy...

Relationship Analysis

Both papers belong to the Mathematical and Code Reasoning category, applying SFT-RL synergy to domains with verifiable rewards. While AceReason-Nemotron focuses on mathematical problem-solving and coding tasks using GRPO with stage-wise RL training on math and code prompts, G1 applies RL to graph-theoretic reasoning tasks using the Erdős dataset. The key difference is the application domain: AceReason targets traditional math/code benchmarks (AIME, LiveCodeBench), whereas G1 addresses graph structure reasoning problems with NetworkX-based verification.

Contributions Analysis

Overall novelty summary. This paper investigates how supervised fine-tuning (SFT) and reinforcement learning (RL) interact to produce strong reasoning models, focusing on data scaling strategies and temperature selection during RL training. It resides in the Mathematical and Code Reasoning leaf, which contains six papers addressing SFT-RL synergy in structured reasoning domains. This leaf sits within the broader Application Domains branch, indicating a moderately populated research direction where domain-specific methods are actively explored. The paper's emphasis on systematic scaling and temperature tuning positions it alongside works examining training orchestration and data quality in mathematical reasoning tasks.

The taxonomy reveals that Mathematical and Code Reasoning is one of several application-focused branches, with neighboring leaves covering Vision-Language Reasoning (16 papers across four sub-leaves) and Specialized Domain Applications (3 papers). The Integration Frameworks branch (12 papers across three leaves) explores general training paradigms, while Theoretical Foundations (9 papers across three leaves) examines mechanistic analyses and comparative studies. The paper's focus on practical training guidelines connects it to Sequential Training Strategies and Mechanistic Analysis leaves, though it remains grounded in mathematical reasoning applications rather than proposing domain-agnostic frameworks.

Among 29 candidates examined, no contributions were clearly refuted. The first contribution (SFT data scaling strategies) examined 10 candidates with zero refutable matches, suggesting limited prior work on systematic prompt versus response scaling comparisons. The second contribution (SFT-RL synergy and temperature selection) examined 9 candidates with no refutations, indicating that the specific temperature-entropy guideline (maintaining 0.3 entropy) may represent a novel empirical finding within the limited search scope. The third contribution (AceReason-Nemotron model) examined 10 candidates with no overlaps, though model releases are inherently unique artifacts. These statistics reflect a top-K semantic search, not exhaustive coverage.

Based on the limited literature search, the work appears to offer incremental advances in understanding SFT-RL interactions for mathematical reasoning. The temperature selection guideline and scaling analysis provide practical insights, though the absence of refutations may partly reflect the search scope rather than absolute novelty. The taxonomy context shows this is an active but not overcrowded research direction, with room for empirical studies that bridge training paradigms and domain-specific applications.

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

Contribution 1: Systematic investigation of SFT data scaling strategies

Description: The authors systematically explore two axes for scaling supervised fine-tuning data: increasing the number of unique prompts and increasing the number of responses per prompt. They find that scaling prompts yields more substantial gains than scaling responses per prompt, and observe consistent performance improvements across training epochs.

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. SFTMix: Elevating Language Model Instruction Tuning with Mixup Recipe

URL: View paper

Brief Assessment

SFTMix[65] focuses on a mixup-based regularization technique for instruction tuning rather than investigating data scaling strategies along the axes of prompt quantity and responses per prompt.

---

2. Matching tasks to objectives: Fine-tuning and prompt-tuning strategies for encoder-decoder pre-trained language models

URL: View paper

Brief Assessment

Matching Tasks Objectives[62] focuses on fine-tuning encoder-decoder models (T5) for specific task-objective alignment with limited epochs (3 epochs), not on systematic scaling of SFT data along multiple axes (prompts vs. responses per prompt) for reasoning models.

---

3. Labeling supervised fine-tuning data with the scaling law

URL: View paper

Brief Assessment

Labeling Scaling Law[67] focuses on manual annotation quality for coreference resolution in chat data, not on systematic scaling strategies for SFT data. The candidate explores annotation calibrated by scaling law principles rather than investigating prompt vs. response scaling strategies or multi-epoch training dynamics.

---

4. Don't Stop Pretraining? Make Prompt-based Fine-tuning Powerful Learner

URL: View paper

Brief Assessment

Prompt-based Powerful Learner[64] focuses on prompt-based continued pre-training for fine-tuning in NLP classification tasks, not on scaling supervised fine-tuning data for reasoning models through prompts vs. responses analysis.

---

5. SLearnLLM: A Self-Learning Framework for Efficient Domain-Specific Adaptation of Large Language Models

URL: View paper

Brief Assessment

SLearnLLM[69] focuses on identifying and fine-tuning only on 'unknown knowledge' (incorrectly answered questions) rather than systematically exploring data scaling strategies across prompts and responses. The candidate does not investigate scaling axes or compare different scaling approaches as the original paper does.

---

6. Codeplan: Unlocking reasoning potential in large language models by scaling code-form planning

URL: View paper

Brief Assessment

CodePlan[66] focuses on scaling code-form planning data for reasoning tasks, not on systematically investigating SFT data scaling strategies along the axes of unique prompts versus responses per prompt as explored in the original paper.

---

7. Entropic distribution matching for supervised fine-tuning of LLMs: Less overfitting and better diversity

URL: View paper

Brief Assessment

Entropic Distribution Matching[60] focuses on addressing overfitting and limited diversity in SFT through entropy regularization and reverse KL divergence, not on systematically exploring data scaling strategies across prompts and responses.

---

8. The best instruction-tuning data are those that fit

URL: View paper

Brief Assessment

Best Instruction-Tuning Data[61] focuses on selecting responses that match the base model's distribution rather than scaling strategies. The candidate does not investigate scaling prompts vs. responses per prompt or multi-epoch training dynamics.

---

9. Frontier AI From the Outside In: Advances in Data Curation, Data Distillation and Model Evaluation

URL: View paper

Brief Assessment

Frontier AI Outside In[68] briefly mentions SFT data scaling and evaluation frequency but does not systematically investigate the two specific axes (scaling prompts vs. responses per prompt) or multi-epoch training dynamics that the original paper explores in depth.

---

10. A Self-Supervised Reinforcement Learning Approach for Fine-Tuning Large Language Models Using Cross-Attention Signals

URL: View paper

Brief Assessment

Self-Supervised Cross-Attention[63] focuses on reinforcement learning using cross-attention signals for self-supervised rewards, not on supervised fine-tuning data scaling strategies involving prompts and responses.

Contribution 2: Analysis of SFT-RL synergy and temperature selection guideline

Description: The authors investigate how different SFT initializations affect final RL performance and establish that stronger SFT models lead to better outcomes when RL is conducted effectively. They provide a rule of thumb for setting sampling temperature to maintain temperature-adjusted entropy around 0.3 for effective exploration-exploitation balance.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. â¦ -Technical Framework for Engagement-Optimized Short Text Generation in Digital Commerce Using Large Language Models and Reinforcement Learning

URL: View paper

Brief Assessment

Engagement-Optimized Text Generation[72] focuses on short text generation for digital commerce using temperature-based sampling as a hyperparameter, not on investigating SFT-RL synergy or establishing temperature selection guidelines for exploration-exploitation balance in reasoning tasks.

---

2. Probing the Origins of Reasoning Performance: Representational Quality for Mathematical Problem-Solving in RL vs SFT Finetuned Models

URL: View paper

Brief Assessment

Probing Reasoning Origins[74] focuses on mechanistic interpretability of RL vs SFT models through linear probing and ablation studies, not on SFT initialization effects or temperature selection for RL training. The candidate does not address the synergy between SFT and RL training dynamics or provide guidelines for sampling temperature selection.

---

3. A Llama walks into the 'Bar': Efficient Supervised Fine-Tuning for Legal Reasoning in the Multi-state Bar Exam

URL: View paper

Brief Assessment

Llama Bar Exam[73] focuses exclusively on supervised fine-tuning for legal reasoning tasks without any reinforcement learning component. The paper does not investigate SFT-RL synergy, temperature effects on RL training, or exploration-exploitation tradeoffs in RL contexts.

---

4. Thyme: Think Beyond Images

URL: View paper

Brief Assessment

Thyme[75] focuses on multimodal vision-language models with code generation for image manipulation tasks, not on general reinforcement learning frameworks for math/code reasoning or SFT initialization effects on RL performance.

---

5. Beyond Two-Stage Training: Integrating SFT and RL for Improved Reasoning in LLMs

URL: View paper

Brief Assessment

Beyond Two-Stage Training[26] focuses on bilevel optimization to integrate SFT and RL simultaneously during training, rather than analyzing how different SFT initializations affect subsequent RL performance or providing temperature selection guidelines for RL

training. The candidate's core contribution is a novel training framework (BRIDGE) that uses nested optimization, not an empirical analysis of SFT-RL dynamics or temperature tuning rules.

---

6. Exploration Strategies for Reasoning Fine-tuning

URL: View paper

Brief Assessment

Exploration Strategies Reasoning[76] focuses on exploration strategies (bottom_p sampling, adaptive temperature) during RL training on a 0.5B model for mathematical puzzles, not on analyzing SFT-RL synergy or establishing temperature guidelines for maintaining entropy around 0.3 across different SFT initializations.

---

7. REINFORCEMENT LEARNING

URL: View paper

Brief Assessment

Reinforcement Learning[77] focuses on multi-turn agentic RL in interactive environments (TextWorld, AlfWorld, SWE-Gym), examining environment complexity, policy algorithms, and reward density. It does not investigate SFT initialization effects on RL performance or provide temperature-adjusted entropy guidelines for exploration-exploitation balance in reasoning models.

---

8. MT: Scaling MLLM-based Text Image Machine Translation via Multi-Task Reinforcement Learning

URL: View paper

Brief Assessment

MT Multi-Task RL[71] focuses on text image machine translation using multi-task RL for recognition, reasoning, and translation tasks. It does not investigate SFT initialization effects on RL performance or provide temperature selection guidelines for exploration-exploitation balance in general RL training.

---

9. Reinforcement Learning with Supervised Alignment

URL: View paper

Brief Assessment

RL Supervised Alignment[70] focuses on constructing supervised alignment for reward modeling in QA tasks, not on analyzing SFT initialization effects or temperature-adjusted entropy guidelines for RL training.

Contribution 3: AceReason-Nemotron-1.1 7B model achieving state-of-the-art performance

Description: The authors develop AceReason-Nemotron-1.1, a 7B parameter model that combines their strong SFT foundation with stage-wise RL training. The model achieves new state-of-the-art results among Qwen2.5-7B-based models on math and code benchmarks, validating their integrated post-training approach.

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. Phi-4-Mini-Reasoning: Exploring the Limits of Small Reasoning Language Models in Math

URL: View paper

Brief Assessment

Phi-4 Mini Reasoning[56] focuses on a 3.8B parameter model using a different training recipe (distillation mid-training, SFT, rollout DPO, RL) rather than the stage-wise RL approach on math-only and code-only prompts used in AceReason-Nemotron-1.1. The candidate does not challenge the novelty of the original's specific model or training methodology.

---

2. Can Compact Language Models Search Like Agents? Distillation-Guided Policy Optimization for Preserving Agentic RAG Capabilities

URL: View paper

Brief Assessment

Compact Agents Distillation[59] focuses on distillation-guided policy optimization for compact models (0.5-1B parameters) in agentic RAG tasks, not on developing 7B reasoning models through SFT-RL synergy for math and code benchmarks.

---

3. Improving Post-Training Quantization via Probabilistic Programming

URL: View paper

Brief Assessment

Post-Training Quantization[55] focuses on neural network quantization techniques for edge device deployment, not on developing reasoning models for math and code tasks. The technical domains are entirely different.

---

4. Efficient inference for large reasoning models: A survey

URL: View paper

Brief Assessment

Efficient Inference Survey[53] is a survey paper that reviews efficient inference methods for large reasoning models. It does not present a specific 7B model or claim to develop AceReason-Nemotron-1.1, thus cannot refute the novelty of this specific model contribution.

---

5. Model compression and efficient inference for large language models: A survey

URL: View paper

Brief Assessment

Model Compression Survey[51] is a comprehensive survey on compression techniques (quantization, pruning, distillation) for large language models. It does not present a specific 7B reasoning model or claim to achieve state-of-the-art results on math and code benchmarks, thus cannot refute the novelty of AceReason-Nemotron-1.1's specific model development and performance claims.

---

6. KV Cache Transform Coding for Compact Storage in LLM Inference

URL: View paper

Brief Assessment

KV Cache Transform[58] focuses on KV cache compression for LLM inference efficiency, not on developing reasoning models through SFT and RL training methods.

---

7. Reinforcement Learning Fine-Tuning of Language Model for Instruction Following and Math Reasoning

URL: View paper

Brief Assessment

RL Fine-Tuning Instructions[50] focuses on a 0.5B parameter model (Qwen2.5-0.5B base) for instruction following and math reasoning, while the original contribution describes a 7B parameter model. The candidate does not demonstrate prior work on developing state-of-the-art 7B models combining SFT and stage-wise RL training.

---

8. To code or not to code? adaptive tool integration for math language models via expectation-maximization

URL: View paper

Brief Assessment

Adaptive Tool Integration[52] focuses on autonomous code integration for math problems via expectation-maximization, not on general SFT-RL synergy or stage-wise RL training methods. The candidate addresses a different technical problem (metacognitive tool-use decisions) than the original's post-training recipe combining SFT scaling with stage-wise RL.

---

9. LightReasoner: Can Small Language Models Teach Large Language Models Reasoning?

URL: View paper

Brief Assessment

LightReasoner[54] focuses on a fundamentally different approach: using smaller language models to teach larger ones through contrastive supervision and expert-amateur divergence, rather than developing a specific 7B model through SFT-RL synergy for math and code reasoning.

---

10. THINKSLM: Towards Reasoning in Small Language Models

URL: View paper

Brief Assessment

THINKSLM[57] is a benchmarking study evaluating existing small language models' reasoning capabilities, not a model development paper. It does not present a competing 7B model or challenge the novelty of AceReason-Nemotron-1.1's post-training approach.

Appendix: Text Similarity Detection

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

References

- [0] AceReason-Nemotron 1.1: Advancing Math and Code Reasoning through SFT and RL Synergy View paper
- [1] Visual-rft: Visual reinforcement fine-tuning View paper
- [2] On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting View paper
- [3] Step-wise Adaptive Integration of Supervised Fine-tuning and Reinforcement Learning for Task-Specific LLMs View paper
- [4] Reason-rft: Reinforcement fine-tuning for visual reasoning View paper
- [5] Teaching large language models to reason with reinforcement learning View paper
- [6] Demystifying long chain-of-thought reasoning in llms View paper
- [7] SRFT: A Single-Stage Method with Supervised and Reinforcement Fine-Tuning for Reasoning View paper
- [8] Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models View paper
- [9] Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering View paper
- [10] Bridging Supervised Learning and Reinforcement Learning in Math Reasoning View paper
- [11] UFT: Unifying Supervised and Reinforcement Fine-Tuning View paper
- [12] Fine-tuning large vision-language models as decision-making agents via reinforcement learning View paper
- [13] Reft: Reasoning with reinforced fine-tuning View paper
- [14] Reasoning with reinforced functional token tuning View paper
- [15] Learning What Reinforcement Learning Can't: Interleaved Online Fine-Tuning for Hardest Questions View paper
- [16] ERank: Fusing Supervised Fine-Tuning and Reinforcement Learning for Effective and Efficient Text Reranking View paper
- [17] Unlock the Correlation between Supervised Fine-Tuning and Reinforcement Learning in Training Code Large Language Models View paper
- [18] Reasoning beyond limits: Advances and open problems for llms View paper
- [19] Advancing reasoning in large language models: Promising methods and approaches View paper
- [20] SuperRL: Reinforcement Learning with Supervision to Boost Language Model Reasoning View paper
- [21] Enhancing LLMs' Reasoning-Intensive Multimedia Search Capabilities through Fine-Tuning and Reinforcement Learning View paper
- [22] Process-Supervised Reinforcement Learning for Interactive Multimodal Tool-Use Agents View paper
- [23] Advancing Multimodal Reasoning via Reinforcement Learning with Cold Start View paper
- [24] Reasoning-Table: Exploring Reinforcement Learning for Table Reasoning View paper
- [25] Blending Supervised and Reinforcement Fine-Tuning with Prefix Sampling View paper
- [26] Beyond Two-Stage Training: Integrating SFT and RL for Improved Reasoning in LLMs View paper
- [27] Grounded Reinforcement Learning for Visual Reasoning View paper
- [28] G1: Teaching LLMs to Reason on Graphs with Reinforcement Learning View paper
- [29] Reassessing the Role of Supervised Fine-Tuning: An Empirical Study in VLM Reasoning View paper
- [30] Anchored Supervised Fine-Tuning View paper
- [31] ViSurf: Visual Supervised-and-Reinforcement Fine-Tuning for Large Vision-and-Language Models View paper
- [32] UAV-VL-R1: Generalizing Vision-Language Models via Supervised Fine-Tuning and Multi-Stage GRPO for UAV Visual Reasoning View paper
- [33] The Synergy Dilemma of Long-CoT SFT and RL: Investigating Post-Training Techniques for Reasoning VLMs View paper
- [34] Reasoning through Exploration: A Reinforcement Learning Framework for Robust Function Calling View paper
- [35] MIRG-RL: Multi-Image Reasoning and Grounding with Reinforcement Learning View paper
- [36] Effective Reinforcement Learning for Reasoning in Language Models View paper
- [37] Beyond Two-Stage Training: Cooperative SFT and RL for LLM Reasoning View paper
- [38] Omni-AutoThink: Adaptive Multimodal Reasoning via Reinforcement Learning View paper

- [39] Token-Efficient RL for LLM Reasoning View paper
- [40] RL Fine-Tuning Heals OOD Forgetting in SFT View paper
- [41] BREAD: Branched Rollouts from Expert Anchors Bridge SFT & RL for Reasoning View paper
- [42] ARES: Alternating Reinforcement Learning and Supervised Fine-Tuning for Enhanced Multi-Modal Chain-of-Thought Reasoning Through Diverse AI Feedback View paper
- [43] How Much Backtracking is Enough? Exploring the Interplay of SFT and RL in Enhancing LLM Reasoning View paper
- [44] Offline RL by Reward-Weighted Fine-Tuning for Conversation Optimization View paper
- [45] On the Impact of Fine-Tuning on Chain-of-Thought Reasoning View paper
- [46] RL Squeezes, SFT Expands: A Comparative Study of Reasoning LLMs View paper
- [47] SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models View paper
- [48] Scaling RL to Long Videos View paper
- [49] Metis-RISE: RL Incentivizes and SFT Enhances Multimodal Reasoning Model Learning View paper
- [50] Reinforcement Learning Fine-Tuning of Language Model for Instruction Following and Math Reasoning View paper
- [51] Model compression and efficient inference for large language models: A survey View paper
- [52] To code or not to code? adaptive tool integration for math language models via expectation-maximization View paper
- [53] Efficient inference for large reasoning models: A survey View paper
- [54] LightReasoner: Can Small Language Models Teach Large Language Models Reasoning? View paper
- [55] Improving Post-Training Quantization via Probabilistic Programming View paper
- [56] Phi-4-Mini-Reasoning: Exploring the Limits of Small Reasoning Language Models in Math View paper
- [57] THINKSLM: Towards Reasoning in Small Language Models View paper
- [58] KV Cache Transform Coding for Compact Storage in LLM Inference View paper
- [59] Can Compact Language Models Search Like Agents? Distillation-Guided Policy Optimization for Preserving Agentic RAG Capabilities View paper
- [60] Entropic distribution matching for supervised fine-tuning of LLMs: Less overfitting and better diversity View paper
- [61] The best instruction-tuning data are those that fit View paper
- [62] Matching tasks to objectives: Fine-tuning and prompt-tuning strategies for encoder-decoder pre-trained language models View paper
- [63] A Self-Supervised Reinforcement Learning Approach for Fine-Tuning Large Language Models Using Cross-Attention Signals View paper
- [64] Don't Stop Pretraining? Make Prompt-based Fine-tuning Powerful Learner View paper
- [65] SFTMix: Elevating Language Model Instruction Tuning with Mixup Recipe View paper
- [66] Codeplan: Unlocking reasoning potential in large language models by scaling code-form planning View paper
- [67] Labeling supervised fine-tuning data with the scaling law View paper
- [68] Frontier AI From the Outside In: Advances in Data Curation, Data Distillation and Model Evaluation View paper
- [69] SLearnLLM: A Self-Learning Framework for Efficient Domain-Specific Adaptation of Large Language Models View paper
- [70] Reinforcement Learning with Supervised Alignment View paper
- [71] MT: Scaling MLLM-based Text Image Machine Translation via Multi-Task Reinforcement Learning View paper
- [72] â¦ -Technical Framework for Engagement-Optimized Short Text Generation in Digital Commerce Using Large Language Models and Reinforcement Learning View paper
- [73] A Llama walks into the 'Bar': Efficient Supervised Fine-Tuning for Legal Reasoning in the Multi-state Bar Exam View paper
- [74] Probing the Origins of Reasoning Performance: Representational Quality for Mathematical Problem-Solving in RL vs SFT Finetuned Models View paper
- [75] Thyme: Think Beyond Images View paper
- [76] Exploration Strategies for Reasoning Fine-tuning View paper
- [77] REINFORCEMENT LEARNING View paper