EVOR Knowledge Base Evolution and LLM Reasoning Performance in MultiPL-E

Assignee Research

June 12, 2026

Abstract

Large Language Models (LLM) with reasoning capabilities offer a promising path for improving candidate evaluation in planning frameworks, but their relative performance against traditional non-reasoning models remains largely underexplored. In this study, we benchmark a distilled 1.5B parameter reasoning model (DeepSeek-R1) against several state-of-the-art non-reasoning LLMs within a generator-discriminator LLM planning framework for the text-to-SQL task. For this, we introduce a novel method for extracting soft scores from the chain-of-thought (CoT) outputs from reasoning that enables fine-gr

1 Introduction

This paper examines: When Reasoning Beats Scale: A 1.5B Reasoning Model Outranks 13B LLMs as Discriminator. Research question: What is the impact of EVOR's diverse knowledge base evolution on the reasoning capability of LLMs as measured by functional correctness scores on the MultiPL-E Python and Java subsets?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.3/10.

3 Results

13 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 8.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
DeepSeek-R1 achieves 87% higher F1 as well as 3.7% better discrimination accuracy than CodeLlama-7B.	✓	0.28
DeepSeek-R1 achieves 3.7% higher execution accuracy than CodeLlama-13B.	✓	0.21
Logit-based soft scoring vs. binary true/false discrimination yields minimal differences ($<$ 1.5%).	✓	0.29
Adding more context or compute budget yields diminishing returns (e.g., $<$0.4% gain beyond 1024 tokens).	✓	0.24
Extremely low budgets severely degrade performance ($<$ 2% accuracy, $>$ 94% failure).	✓	0.21
DeepSeek-R1 underperforms as a generator, even compared to smaller non-reasoning LLMs.	✓	0.21
Generation is more challenging than discrimination for reasoning models.	✓	0.17
The framework adopts a generator-discriminator approach where the generator LLM proposes candidate solutions and the dis	✓	0.16
The planning module ranks candidates based on the evaluation outcomes and orchestrates the interaction between the two m	✓	0.20
The framework includes an environment component that checks the executability of the candidates.	×	0.07

References

- http://arxiv.org/abs/2505.03786v1
- http://arxiv.org/abs/2601.01982v1

- http://arxiv.org/abs/2503.15113v1