Comparative Analysis of Reasoning Accuracy in Multimodal Large Language Models and Diffusion-Based Trajectory Policies on

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: How does the reasoning accuracy of multimodal large language models compare to diffusion-based trajectory policies in dynamic task planning environments when evaluated on the RoboBench benchmark with varying levels of environmental noise?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.6/10.

3 Results

16 papers retrieved. 10 claims extracted; 8 independently verified. Quality review score: 7.6/10.

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
A 1.5B distilled DeepSeek-R1 model achieves 87% higher F1 score than CodeLlama-7B in discrimination tasks.	✓	0.20
A 1.5B distilled DeepSeek-R1 model achieves 3.7% better discrimination accuracy than CodeLlama-7B.	✓	0.23
A 1.5B distilled DeepSeek-R1 model achieves 3.7% higher execution accuracy than CodeLlama-13B.	✓	0.23
Using logit-based soft scoring versus binary true/false discrimination yields performance differences of less than 1.5%.	✓	0.20
Increasing compute budget beyond 1024 tokens yields less than 0.4% performance gain for reasoning models.	×	0.13
Using extremely low compute budgets results in less than 2% accuracy and greater than 94% failure rate for reasoning mod	×	0.11
DeepSeek-R1 underperforms as a generator compared to smaller non-reasoning LLMs.	✓	0.21
The study evaluates LLMs' ability to discriminate correct and incorrect SQL queries by relabeling oracle-generated outp	✓	0.24
Discriminator performance was tested in a naive setting and an enhanced setting that filters by executability via enviro	✓	0.25
Traditional non-reasoning LLMs can provide logit-based scores for discrimination, whereas reasoning models produce arbit	✓	0.23

References

- http://arxiv.org/abs/2505.03786v1
- http://arxiv.org/abs/2404.07214v4
- http://arxiv.org/abs/2410.05821v2