How does domain-specific fine-tuning affect the performance of tabular foundation models on structured reasoni

Assignee Research

June 10, 2026

Abstract

We study a pipeline that curates reasoning data from initial structured data for improving long-context reasoning in large language models (LLMs). Our approach, $$\pi^2, constructs high-quality reasoning data through rigorous QA curation: 1) extracting and expanding tables from Wikipedia, 2) from the collected tables and relevant context, generating realistic and multi-hop analytical reasoning questions whose answers are automatically determined and verified through dual-path code execution, and 3) back-translating step-by-step structured reasoning traces as solutions of QA pairs given realistic

1 Introduction

This paper examines: $$\pi^2: Structure-Originated Reasoning Data Improves Long-Context Reasoning Ability of Large Language Models. Research question: How does domain-specific fine-tuning affect the performance of tabular foundation models on structured reasoning benchmarks like TabMWP or TabFact, measured by accuracy and robustness under distribution shifts?.

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 4.1/10.

3 Results

16 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 4.1/10.

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
LongSeal is a multi-document QA benchmark containing 254 questions, where each question includes one ground-truth docume	×	0.03
The LongBenchV2 Medium subset used in the experiment contains 215 samples with context lengths ranging from 32k to 128k	×	0.02
The Oolong benchmark evaluation used 400 samples from the Oolong-Synth test set with a context length of 64k tokens.	×	0.05
The OfficeQA dataset contains 246 questions based on U.S. Treasury Bulletin PDFs from 1939 to 2025.	×	0.02
GEMINI 3.1 FLASH-LITE achieved an average score of 59.58 across the five evaluated benchmarks.	×	0.03
SFT (π2-20B-A3B) with high reasoning effort achieved a score of 64.0 on the LongSeal benchmark.	×	0.05
The base QWEN3-4B-INSTRUCT-2507 model achieved a score of 14.88 on the OfficeQA benchmark.	×	0.08
Supervised finetuning (SFT) was performed using LoRA with a rank of 8 on 922 training samples from π2.	×	0.02
The SFT process for each model took 2.5 hours on 4 NVIDIA H200 GPUs.	×	0.02
The overall improvements brought by SFT on π2 are statistically significant at α = 0.01.	×	0.01
GPT-OSS-120B achieved a score of 75.0 on the π2-Bench.	×	0.09
SFT (π2-4B) achieved a score of 40.94 on the LongSeal benchmark, compared to 33.07 for its base model.	×	0.01

References

- http://arxiv.org/abs/2512.03307v1
- http://arxiv.org/abs/2604.05114v1

- http://arxiv.org/abs/2601.04110v2