arxiv.org/abs/2508.15754

# Dissecting Tool-Integrated Reasoning: An Empirical Study and Analysis

**Yufeng Zhao**[1,2,*], **Junnan Liu**[1,*], **Hongwei Liu**[1], **Dongsheng Zhu**[1], **Yuan Shen**[2,†], **Songyang Zhang**[1,†], **Kai Chen**[1,†]

[1]Shanghai Artificial Intelligence Laboratory
[2]Department of Electronic Engineering, Tsinghua University
{zhaoyufeng,zhangsongyang}@pjlab.org.cn, to.liujn@outlook.com

ReasonZoo

## 简介

本文提出ReasonZoo，涵盖9类任务的TIR评测基准，此外，为了评估TIR的推理效率(用更少的token完成推理)，作者提出了两个新的指标PAC和AUC-PCC。不啰嗦了，直接说实验结论吧：

- TIR模型整体优于非TIR 模型，无论在数学还是非数学领域
- lrm size越大，提升越明显，尤其再配合更复杂的TIR范式
- 经过PAC和AUC-PCC分析，TIR是能够减少冗余推理的，也就是缓解overthinking问题
- TIR的增益既来自外部工具反馈，也体现在lrm的thinking优化，但是也存在推理被外部工具带偏反而得到错误答案的情况

## 背景

本文属于TIR (tool-integrated reasoning)方向并且是benchmark类型的工作。作者想弄清楚三个问题：

1) TIR能提升lrm的数学推理基本没有疑义，这种效果提升能否泛化到其他领域呢?
2) TIR能带来效果提升到底是因为借助外部tool(获取答案/外部数据)还是提升了lrm本身的推理能力呢?
3) TIR是否也能削弱overthinking，或者说提升推理效率呢?

## 实验设置

- 做实验的两类模型：lrm基座(Qwen3和DeepSeek-R1-0528)和专门优化做TIR的模型(CIR和ToRL)
- 推理效率：

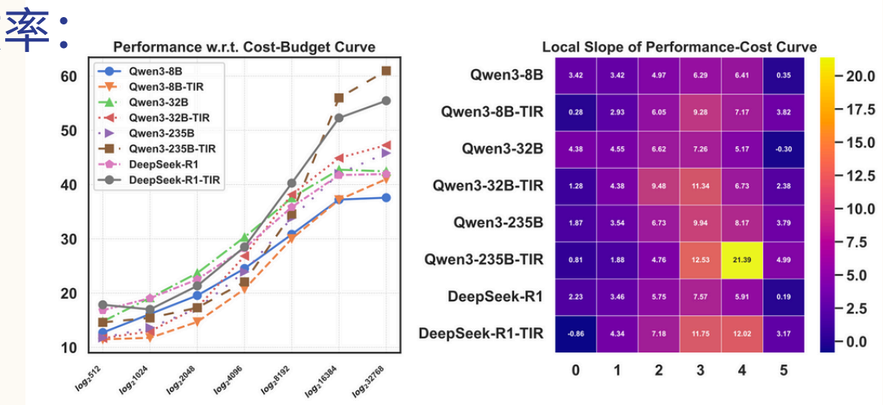
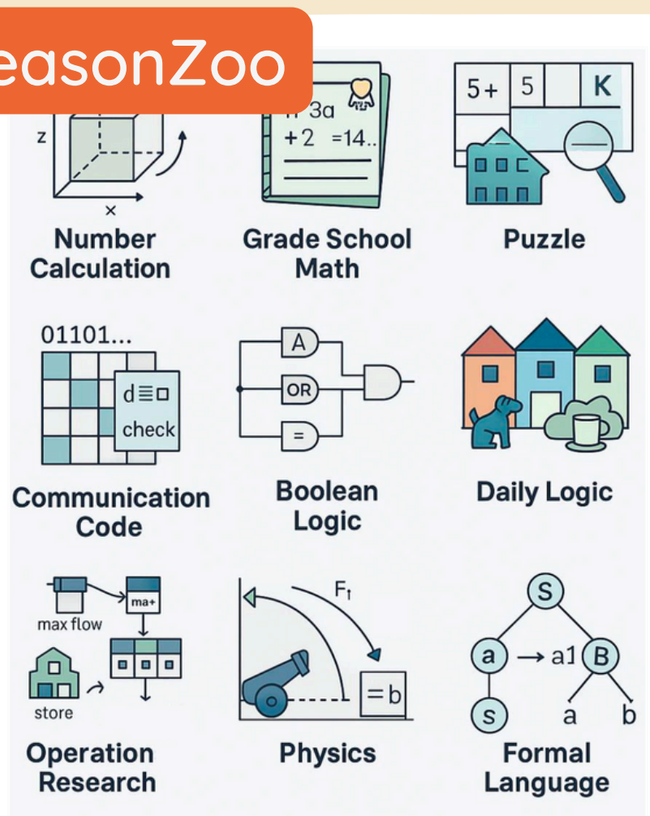


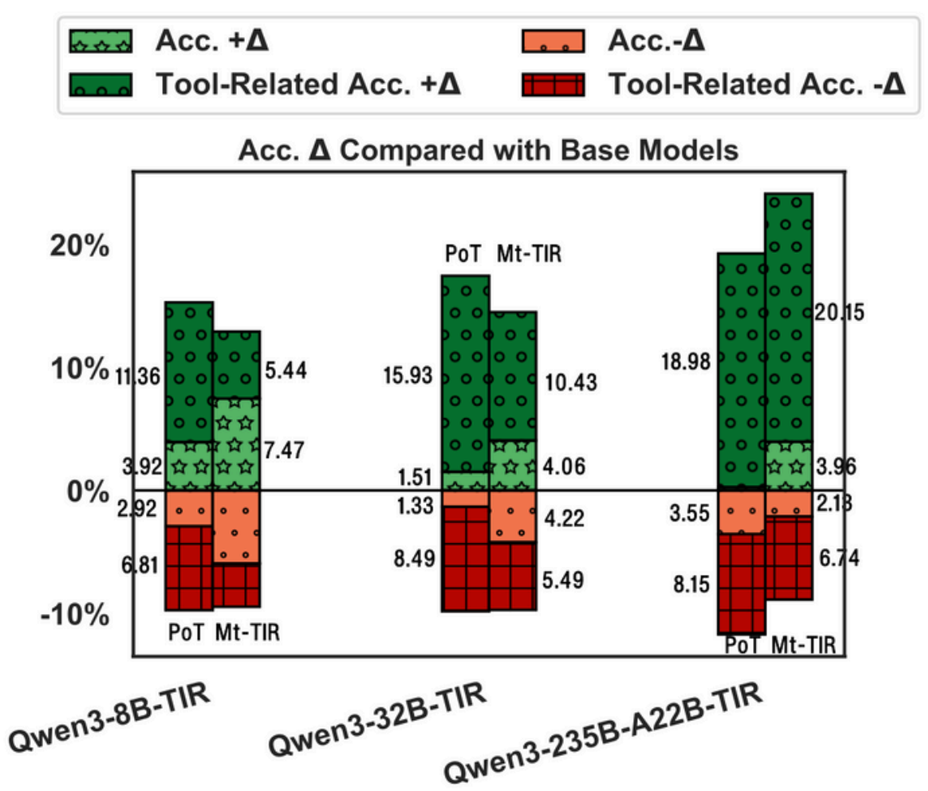

Figure 2: Illustration of Performance w.r.t. Cost-Budget Curve on REASONZOO.

## ReasonZoo




我比较感兴趣对TIR效果提升的归因分析，见右图，作者找出base model和TIR模型答案不一致的题目，再用Qwen2.5-32B-Instruct来分析差异究竟是否依赖于工具反馈，从而把TIR增益分解为Tool-Related Acc. +Δ 与 Acc. +Δ 两部分。

- 工具反馈带来的效果提升非常明显
- lrm本身推理能力得到提升不太明显
- 工具导致的错误也不少




## 思考

通过实验来看TIR是能切切实实提升推理效果的，并且推理效率也有所提升，从归因分析来看，如何减少工具带来的错误或者更好的将工具反馈融入lrm推理轨迹是一个非常值得思考的问题，我们之前也读过几篇相关的论文，比如ARPO(基于GRPO的熵感知TIR策略优化方法)、AutoRefine(在search + TIR推理时，让llm对检索结果先精炼再推理)，对这个问题感兴趣的朋友可以去回顾下。