

思维定理：一种用于语言模型中溯因、演绎和归纳推理的多代理框架

Samir Abdaljalil*, Hasan Kurban, Khalid Qaraqe, Erchin Serpedin*

*Texas A & M University, College Station, TX., USA

Hamad Bin Khalifa University, Doha, Qatar

sabdaljalil@tamu.edu, hkurban@hbku.edu.qa

Abstract

Large language models (LLMs) have shown strong performance across natural language reasoning tasks, yet their reasoning processes remain brittle and difficult to interpret. Prompting techniques like Chain-of-Thought (CoT) enhance reliability by eliciting intermediate reasoning steps or aggregating multiple outputs. However, they lack mechanisms for enforcing logical structure and assessing internal coherence. We introduce Theorem-of-Thought (ToTh), a novel framework that models reasoning as collaboration among three parallel agents, each simulating a distinct mode of inference: abductive, deductive, and inductive. Each agent produces a reasoning trace, which is structured into a formal reasoning graph. To evaluate consistency, we apply Bayesian belief propagation guided by natural language inference (NLI), assigning confidence scores to each step. The most coherent graph is selected to derive the final answer. Experiments on symbolic (WEBOfLIES) and numerical (MULTIARITH) reasoning benchmarks show that ToTh consistently outperforms CoT, Self-Consistency, and CoT-Decoding across multiple LLMs, while producing interpretable and logically grounded reasoning chains. Our findings suggest a promising direction for building more robust and cognitively inspired LLM reasoning. The implementation is available at <https://github.com/KurbanIntelligenceLab/theorem-of-thought>.

1 引言

大型语言模型 (LLMs) 在各种自然语言理解和生成任务中取得了令人印象深刻的表现，得益于情境学习、指令微调和思维链 (CoT) 提示方面的进展。这些方法扩展了 LLMs 在处理复杂推理形式方面的能力，包括数学、逻辑和常识推理。

尽管这些进展取得了，LLM 的推理仍然浅显且不可靠。现有方法常常依赖于单次或基于采样的线性推理路径的解码，使得这些方法容易受到幻觉 (?)、逻辑不一致 (?) 和弱泛化 (?) 的影响。诸如 CoT 和自一致性 (??) 的方法鼓励中间步骤和对采样输出的多数投票，但缺乏验证内部一致性和模拟推理逻辑结构的机制。因此，输出可能看起来流畅且似乎合理，但在逻辑上仍然不健全。

这种脆弱性与人类推理形成了鲜明对比，人类推理本质上是多方面的。借鉴认知科学的见解 (?)，我们观察到人类推理通常融合了三种互补的模式——溯因、演绎和归纳——支持解释、推导和泛化。然而，大型语言模型通常将这些不同的过程混淆为一个单一且未分化的流程，限制了可解释性和可靠性。

为了解决这一差距，我们提出了思维定理 (ToTh)，这是一个通过结构化、可验证的交互来建模多样化推理策略的框架。ToTh 使用三个专门的代理，每个代理模拟一种独特的认知模式：

- 溯因：为观察到的事实推断出合理的解释；
- 演绎：从给定的前提推导出有效的结论；
- 归纳：从模式或例子中推广。

每个代理独立生成推理轨迹，该轨迹被转换为形式推理图 (FRG) ——一种有向图，其中节点代表中间结论，边捕捉逻辑依赖关系。我们使用贝叶斯置信传播评估每个 FRG 的内部一致性，其中边的置信评分通过自然语言推理 (NLI) 模型进行校准。通过平衡平均信念和逻

辑熵的复合评分来选择最连贯的图形，从中提取最终答案。

贡献 本工作的关键结果是：

- 我们介绍了 ToTh，这是一个结构化推理框架，将溯因推理、演绎推理和归纳推理整合到一个模块化的基于 LLM 的流程中。
- 我们在推理图上开发了一种信念传播机制，利用自然语言推理 (NLI) 通过贝叶斯更新来评估和评分逻辑一致性。
- 我们证明了 ToTh 在多个大型语言模型中始终优于最新的推理方法（例如，CoT、自洽性、CoT 解码）。
- 我们在符号 (WEBOfLIES) 和数值 (MULTIARITH) 基准测试中的评估突出表明，ToTh 在需要多步骤推理的任务中的稳健性——在这些情况下，直接提示通常会失败 (?)。

本文的其余部分组织如下：第 2 节回顾相关工作。第 3 节介绍了 ToTh 框架。第 4 节描述了实验设置，第 5 节分析了获得的结果。第 6 节总结了对 LLMs 中结构化推理的影响和未来研究方向。

2 相关工作

越来越多的研究探索了提示策略以增强大型语言模型的推理能力。CoT 提示 (?) 鼓励模型将问题分解为中间步骤，引导推理沿着线性路径进行。在此基础上，Auto-CoT (?) 通过采样多样性问题并生成相应的推理线索来自动化提示生成，从而减少了人力投入。除了提示生成外，还有一些工作专注于优化提示选择策略。ActivePrompt (?) 通过主动学习识别高不确定性的实例进行标注，提高了数据效率和推理的鲁棒性。更近期的方法将显式结构引入推理过程。Tree-of-Thought (ToT) (?) 支持带有内部评估的多路径探索，而 Graph-of-Thought (GoT) (?) 则将推理构建为图，以更好地建模步骤之间的依赖关系。

针对推理的指令微调。 指令微调和知识蒸馏提供了在不依赖显式提示的情况下在 LLM 中引发推理的替代方法 (???)。虽然有效，这些方法通常需要对带有推理痕迹和 CoT 示例的大规模数据集进行计算密集型的微调，这通常成本高且领域特定。最近的工作探索了更多间接监督策略。例如，? 介绍了代理微调，这利用辅助模型来对比基础 LLM 及其适配变体。

虽然这种方法减少了对直接监督的需求，但它仍然假设可以访问类似 CoT 的输出和预先对齐的推理基准。

3 方法论

ToTh 是一个基于图的推理框架，旨在提高在复杂任务上 LLM 的准确性、可解释性和泛化能力。它将推理分解为三个模块化的代理，每个代理模拟一种经典的推理范式——归纳法、演绎法和归纳法。每个代理产生一个结构化的推理轨迹，并构成一个 FRG。最终的答案通过 NLI 校准的贝叶斯信念传播和复合图评分得出。完整的流程如图 1 所示。

ToTh 在三个方面与先前的推理范式不同：架构、监督和验证。基于提示的方法（如 CoT、ToT、GoT）通过线性或松散结构的轨迹来引发推理，但缺乏强制逻辑一致性的机制。经过指令调整的模型通过在注释轨迹上进行微调来嵌入推理行为，通常需要大型数据集，并且在推理时保持不透明。虽然这两类方法都反映了对结构化多步骤推理的浓厚兴趣，但它们通常在单一或隐式架构内运作，不支持形式的一致性检查。相比之下，ToTh 实例化了独特的认知代理，将它们的输出整合到一个可解释的图中，并通过 NLI 引导的贝叶斯推理明确验证推理的连贯性，从而实现模块化、透明且可验证的推理，超越现有方法的范围。

给定一个自然语言问题 q ，ToTh 部署了三个独立的求解代理，每个代理都与一种特定的经典推理模式对齐：溯因推理、演绎推理和归纳推理。这些范式的形式定义如下。

溯因推理代理 a_1 在给定一组观察 O 和背景知识 K 的情况下，推断出最合理的假设 H ，形式化为：

$$a_1 : \arg \max_H P(H | O, K).$$

演绎推理代理 a_2 从一组前提 $\{P_1, P_2, \dots, P_n\}$ 中推导出合乎逻辑的结论 C ，表示为：

$$a_2 : \{P_1, P_2, \dots, P_n\} \vdash C.$$

归纳推理代理 a_3 从观察到的例子 $\{x_1, x_2, \dots, x_n\}$ 中归纳出一个规则 R ，表达为：

$$a_3 : \{x_1, x_2, \dots, x_n\} \Rightarrow R.$$

每个代理 $a_i \in \{a_1, a_2, a_3\}$ 独立生成一个推理轨迹

$$\mathbf{r}^{(i)} = [r_1^{(i)}, r_2^{(i)}, \dots, r_{s_i}^{(i)}],$$

其中 $r_j^{(i)}$ 表示代理推理过程中的第 j 步。

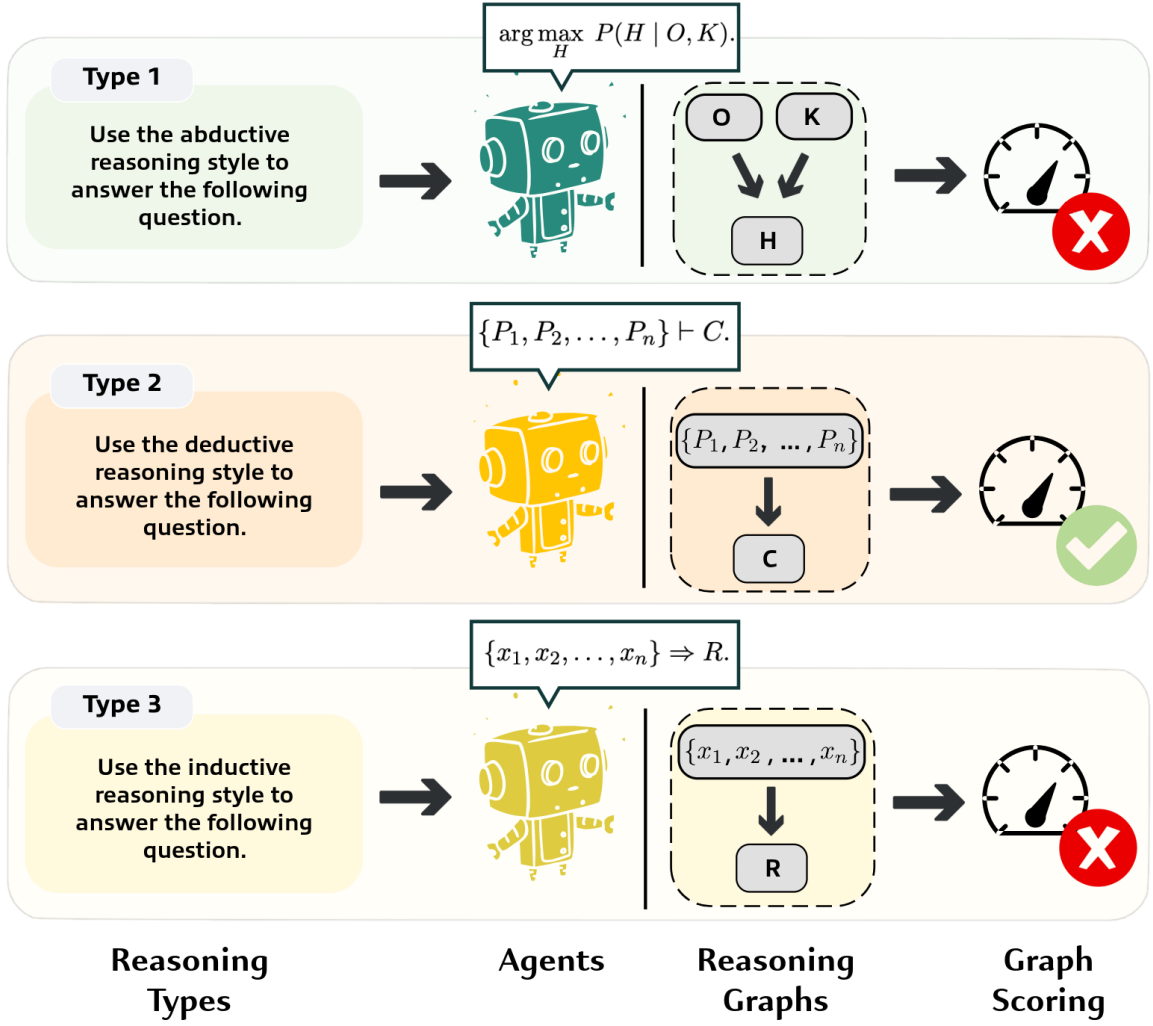


Figure 1: 思维定理 (ToTh) 推理流程概述。一个问题由三个代理独立处理，每个代理使用不同的推理风格：溯因 (1 型)、演绎 (2 型) 和归纳 (3 型)。每个代理生成一个结构化的推理图，该图通过贝叶斯置信传播进行评分。溯因在给定观察 O 和知识 K 的情况下推导出最佳假设 H (即, $\arg \max_H P(H | O, K)$)；演绎从前提 $\{P_1, \dots, P_n\}$ 推导出结论 C (即, $\{P_i\} \vdash C$)；归纳从例子 $\{x_1, \dots, x_n\}$ 推导出规则 R (即, $\{x_i\} \Rightarrow R$)。得分最高的图将其最终节点作为答案。✓ 和 ✗ 表明是否选择了某个给定代理的输出。

每个推理路径 $r^{(i)}$ 被转换为一个有向图 $G^{(i)} = (V^{(i)}, E^{(i)})$ ，其中 $V^{(i)}$ 表示表示单个推理步骤的节点集， $E^{(i)}$ 表示编码这些步骤之间推理关系的有向边。边 $(v_u \rightarrow v_v) \in E^{(i)}$ 是使用一个预训练的 NLI 模型推断出来的，该模型评估推理步骤之间的语义关系。每条边根据预测标签注释一个信任分数 $\theta_{uv} \in [0, 1]$ ：

$$\theta_{uv} = \begin{cases} 0.95 & \text{if entailment} \\ 0.60 & \text{if neutral} \\ 0.10 & \text{if contradiction} \end{cases}$$

这些分数量化了中间步骤之间逻辑蕴涵的强度，为后续的信念传播阶段中的概率推理提供了一个校准的基础。

为了建模图中信念的传播，信念值通过贝叶斯更新规则进行传播，该规则改编自概率图模

型中的传统信念传播公式 (?)。

每个节点 $v \in V$ 都以先验置信度 $P(v) = 0.5$ 初始化，反映了最大的不确定性。对于一个具有单一父节点 v_p 且与信任评分 θ_{pc} 相关联的节点 v_c ，更新后的信念是使用贝叶斯更新规则计算得出的：

$$P(v_c) = \frac{P(v_p) \cdot \theta_{pc}}{P(v_p) \cdot \theta_{pc} + (1 - P(v_p)) \cdot (1 - \theta_{pc})}.$$

对于多个父节点 $\{v_{p_1}, \dots, v_{p_m}\}$ 的情况， v_c 的信念被计算为来自每个父节点的个体更新的平均值：

$$P(v_c) = \frac{1}{m} \sum_{j=1}^m f(P(v_{p_j}), \theta_{p_j c})$$

$$f(p, \theta) = \frac{p \cdot \theta}{p \cdot \theta + (1 - p)(1 - \theta)}.$$

这种递归公式通过图传播信心，在一致的推理路径上放大一致性，而在检测到上游的不确定性或矛盾时减弱信念。

每个推理图 $G^{(i)}$ 的评估基于平均节点置信度和逻辑不确定性之间的权衡。我们优先考虑那些既自信（高置信度）又不确定性较低（低熵）的图。平均置信度计算为

$$\mu^{(i)} = \frac{1}{|V^{(i)}|} \sum_{v \in V^{(i)}} P(v),$$

，标准化二进制熵表示为

$$H^{(i)} = -\frac{1}{|V^{(i)}|} \sum_{v \in V^{(i)}} h(P(v))$$

。最终分数结合了这两个部分：

$$\text{Score}(G^{(i)}) = \mu^{(i)} - H^{(i)}.$$

。得分最高的推理图被选作最终候选：

$$G^* = \arg \max_i \text{Score}(G^{(i)}).$$

。

最终答案是从所选择图 G^* 的终端节点提取的，该节点对应于相关推理路径的最后一步。

理论复杂性 令 $k = 3$ 表示推理代理的数量， s 表示每个代理生成的推理步骤数。ToTh 框架涉及三个主要计算阶段：信任估计、信念传播和图评分。在信任估计阶段，每个代理生成一系列推理步骤，并对每个相邻对应用 NLI 模型来评估逻辑连接的强度。由于每个路径最多包含 $s - 1$ 对，因此所有代理的 NLI 评估总数为 $\mathcal{O}(k \cdot s)$ 。在信念传播阶段，构建的推理图中的每个节点恰好以拓扑顺序访问一次，并根据传入的信任分数使用贝叶斯更新规则更新其后验置信度，结果总共进行了 $\mathcal{O}(k \cdot s)$ 次更新。最后，图评分涉及计算每个图中的所有节点的平均置信度和熵，这也需要 $\mathcal{O}(k \cdot s)$ 时间。因此，ToTh 管道的端到端复杂度是 $\mathcal{O}(k \cdot s)$ ，在线性于代理的数量和每个代理的推理步骤数量。

这使得 ToTh 比自洽性或 CoT 解码等基于采样的方法效率高得多，后者需要 $\mathcal{O}(n)$ 次解码传递，其中 n 是采样推理链的数量。相比之下，ToTh 对每个代理执行单一的结构化推理传递，随后进行轻量级的验证和评分，提供了一种更具可扩展性和可解释性的随机解码替代方案。

4 实验

数据。 ToTh 在两个具有代表性的推理基准上进行了评估。MULTIARITH (?) 针对通过多步骤算术文字题目的组合数值推理。WEBOF LIES (?) 是 BIG-BENCH-HARD 套件的一部分，涉及确定逻辑上纠结的符号陈述之间的真假值。这些数据集在直接提示下被认为对大型语言模型具有挑战性 (?)，使其适合测试结构化推理能力。

为了在规模、对齐和架构上提供多样性，选择了三个公开可用的大语言模型：(1) MISTRAL-7B (?) ¹，一个具有高效扩展的通用解码器模型；(2) DEEPSEEK-7B (?) ²，一个经过指令调优的模型，优化用于多轮推理和对齐；以及 (3) PHI-3.5 MINI (?) ³，一个专为教育、低成本推理任务设计的轻量级模型。此选择范围从紧凑的推理高效模型到指令对齐的推理专注系统。

基线。 ToTh 与三个强大的基线进行了比较：CoT (?)，Self-Consistency (?)，和 CoT-Decoding (?)。CoT 用于提示模型在回答之前生成中间推理步骤。Self-Consistency 通过采样 $n = 20$ 次完成并选择最频繁的答案来提高鲁棒性。CoT-Decoding 通过使用多样化的解码路径刺激潜在的推理行为，消除了显式提示。

所有模型均在其发布形式中进行评估，而未进行微调。解码在温度为 0.7 和最大输出长度为 526 标记的条件下进行。为了评分推理的连贯性，使用了 RoBERTa-MNLI ⁴，符合基于 NLI 的输出验证的先前工作 (?)。所有方法的输入都统一格式化为 “Q: [question] \n A:”，以便与基准保持一致 (?)。

为了指导推理行为，以下指令被添加到每个输入之前，使用每个代理适当的 { style } 关键词：

```
Use the { style } reasoning
style to answer the following
question.
Follow these instructions
carefully:
```

- Break the problem into clear, numbered reasoning

¹<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>

²<https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat>

³<https://huggingface.co/microsoft/Phi-3.5-mini-instruct>

⁴<https://huggingface.co/FacebookAI/roberta-large-mnli>

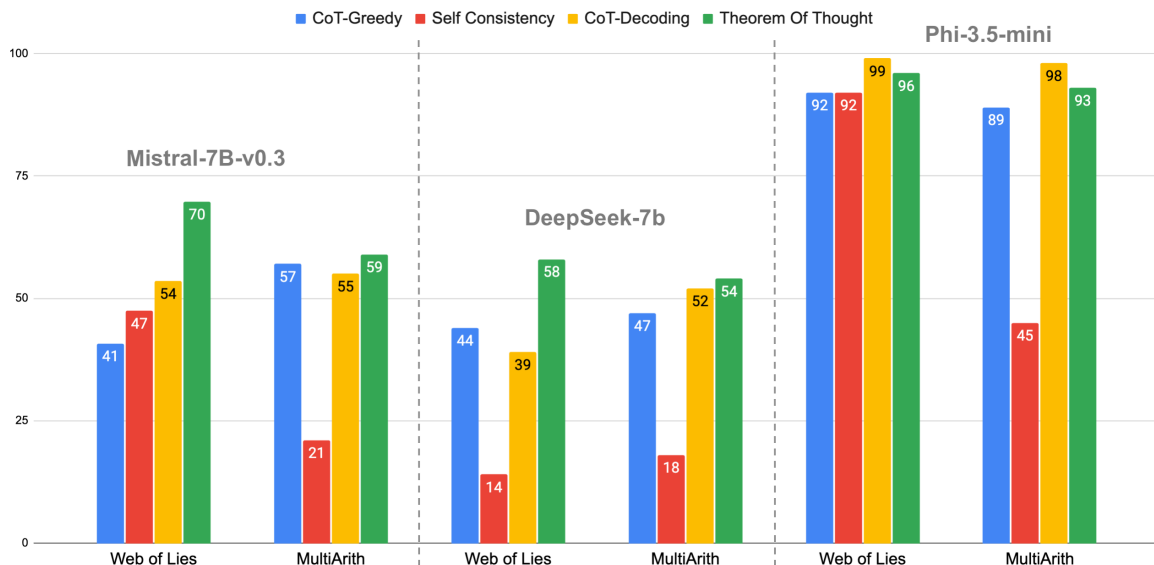


Figure 2: 在两个基准任务（WEBOfLIES 和 MULTiARITH）上使用三种开源语言模型：MISTRAL-7B-v0.3、DEEPSEEK-7B 和 Phi-3.5-mini，对不同推理流程的准确度（%）进行比较。每组条形对应一种不同的推理方法：CoT-Greedy（蓝色）、Self-Consistency（红色）、CoT-Decoding（黄色），以及我们提出的思想定理（绿色）。

steps using { style } .

- Reference any known principles, patterns, or assumptions involved.
- Arrive at a final answer that directly responds to the question.

所有实验对每个输入仅使用一次解码过程。随机种子是固定的，解码设置保持不变以确保可重复性。

5 结果

5.1 主要实验结果

结果以答案准确率（%）的形式报告，并在图 2 中进行了总结。

各模型的性能比较 ToTh 在使用 MISTRAL-7B 和 DEEPSEEK-7B 进行评估时，始终在两个任务上都优于所有基线方法，显示出在推理准确性方面的明显提升。在 PHI-3.5 MINI 上，尽管 CoT-Decoding 在某些实例上略微超越 ToTh，但 ToTh 在象征性和数值任务上始终表现出色。例如，在 WEBOfLIES 数据集上，ToTh 分别在 MISTRAL-7B 和 DEEPSEEK-7B 上比 CoT-Greedy 提高了 29 % 和 14 %，并且在 PHI-3.5 MINI 上与表现最佳的方法相差不超过 3 %。这些结果突出了 ToTh 在不同规模和对齐模型中的鲁棒性和泛化能力。

与 CoT-解码的比较。 虽然 CoT-Decoding 在 Phi-3.5-mini 上表现强劲，几乎在 WEBOfLIES 上取得完美成绩（99 %），但 ToTh 在保持跨模型更高一致性的同时表现相当或略低（96 %）。例如，在 MULTiARITH 数据集上，ToTh 在 MISTRAL-7B 和 DEEPSEEK-7B 上均领先 CoT-Decoding 4-5 分，表明其在数值推理方面具有更强的泛化能力。

令人惊讶的是，自洽性在所有设置中表现不佳，特别是在符号任务上。例如，在有 DEEPSEEK-7B 和 MISTRAL-7B 的条件下，它在 WEBOfLIES 和 MULTiARITH 上仅分别达到 14 % 和 21 %。这表明对于逻辑性强的任务，随机生成的多数投票未能捕获结构化的依赖关系。

正如预期，性能随模型能力而扩展。Phi-3.5-mini 在所有方法中取得了最高的绝对分数，反映了其更强的对齐和训练。然而，即使在较小的模型规模下，ToTh 相比基线的优势仍然有意义，这表明架构对推理稳健性有贡献，而不仅仅是模型规模。尽管 DEEPSEEK-7B 在训练时考虑了推理能力，但其更广泛的训练目标，包括代码生成和开放式问答，可能会分散其在结构化推理任务中的专注。相比之下，Phi-3.5-mini 从专注于教育和逐步问题解决的针对性课程中受益，这可能解释了其在符号和数学基准上的优异表现。有趣的是，尽管 MISTRAL-7B 和 DEEPSEEK-7B 大小相似，但 MISTRAL-7B 始终表现优于 DEEPSEEK-7B。

	WEBOfLIES			MULTIARITH		
	3	4	5	d_0/l_3	d_0/l_4	d_2/l_3
CoT-G	41	32	19	57	26	14
SelfC	48	47	38	21	6	17
CoT-Dec	54	48	46	55	41	24
ToTh	70	56	<u>43</u>	59	45	<u>21</u>

Table 1: 准确性 (%) 在符号 (WEBOfLIES) 和数学 (MULTIARITH) 推理任务中随着难度级别的增加而变化。第 3 到 5 列分别对应具有 3、4 和 5 个相互依赖陈述的符号推理。列 d_0/l_3 、 d_0/l_4 和 d_2/l_3 表示按深度和长度分类的算术推理问题: d 表示运算深度, l 表示序列长度。ToTh 在 6 个设置中 5 次获得最高准确率, 即使在最复杂的实例中也保持竞争力, 在符号和数值领域都表现出一致的性能。加粗: 最佳性能; Underlined: 第二佳。

这可能归因于 Mistral 的更为干净、集中于推理的预训练数据和架构级别的优化, 这增强了其遵循多步骤指令和在标记跨度中保持逻辑连贯性的能力。

5.2 推理复杂度下的鲁棒性

为了评估 ToTh 在推理复杂度增加时的稳健性, 使用 MISTRAL-7B 模型在符号和数字任务上进行了实验。表 1 展示了按问题难度分层的准确性结果: 对于 WEBOfLIES, 按相互依赖语句数 (3-5) 分层; 对于 MULTIARITH, 按操作深度/长度组合分层。

ToTh 在所有难度级别上保持强劲的表现, 超越或接近领先的基准。在符号推理中, ToTh 在最具挑战性的设置 (5 个陈述) 中达到 43 % 的准确率, 显著超越 CoT-Greedy (19 %) 和 Self-Consistency (38 %), 并接近 CoT-Decoding (46 %)。这一趋势在更简单的情况下仍然存在, 其中 ToTh 在 3 和 4 个陈述中取得了最高分。

在数值推理方面, ToTh 在较低复杂度水平上提供了最强的结果——在 d_0/l_3 (59 %) 和 d_0/l_4 (45 %) 实现了最新技术性能——即使在较高复杂度 (d_2/l_3) 时也具有竞争力, 其准确率与 CoT-Decoding 相当 (21 % 对比 24 %)。这些发现突出了 ToTh 在不同任务难度下的泛化能力, 并表明其结构化、多代理推理设计在推理负载增加时提供了可扩展的优势。

6 结论和未来工作

这项工作提出了一种称为 Theorem-of-Thought (ToTh) 的图形推理框架, 通过模块化多代理设计整合了溯因推理、演绎推理和归纳推理。每个代理生成结构化的推理轨迹, 这些轨迹被组合成正式的图并通过自然语言推理标定的贝叶斯置信度传播进行验证。这种方法支持准确的预测和可解释的、逻辑上有依据的

推理。在符号和数值基准上的实证评估表明, ToTh 在需要结构化逻辑推理的场景中持续优于强力提示和解码基线。

ToTh 通过将推理视为一个可验证的、组合的过程, 而不是一个单一的生成任务, 引入了一种在语言模型中推理的新范式。未来的研究将探索基于输入特征的动态代理路由、代理间协作协议, 以及通过微调和基于集合的自然语言推理模型进行自适应信任估计。将这一框架扩展到科学假设验证、法律和政策推理以及视觉问答等多模态领域, 代表了在大型语言模型中推进通用的、可验证推理的一个有前途的方向。

7

局限性

固定推理类型。 ToTh 假定在所有输入中均匀地分解为溯因、演绎和归纳推理。虽然这种模块化提高了可解释性, 但它也施加了一个固定的认知框架, 这可能与需要混合或非典型推理模式的任务不匹配。例如, 创造性任务或模糊提示可能受益于动态地混合推理类型或强调某一类型而非其他。这种僵化性可能限制 ToTh 的适应性, 并在此类情况下导致次优的轨迹组成。未来的工作可能探索数据驱动和上下文敏感的代理路由, 允许框架根据输入语义有选择地实例化和抑制推理模式。

贝叶斯置信度传播机制对低置信度节点中的噪声敏感, 这可能会削弱本应有效的推理链或扭曲图中更深区域的信念估计。这种情况可能发生在较长的推理过程中, 早期推理步骤中的错误不成比例地传播, 从而降低最终预测的可靠性。此外, 目前的传播是均匀且未规范化的, 缺乏针对对抗和不一致中间步骤的稳健性机制。结合经过校准的不确定性建模、边掉落和置信度平滑——可能受细粒度蕴含分布的启发——可以提高稳定性并缓解局部不一致性的放大。