神经符号学习：理论、方法与应用

李昊轩
北京大学人工智能研究院助理教授、研究员、博士生导师
北京中关村学院兼职教授、博士生导师

研究背景



大模型发展进入后训练的新阶段，其中大模型幻觉是当前关键问题

基于scaling-law驱动的预训练：Deepseekv3，GPT4o

基于reasoning能力增强的后训练：GPT-o1，Deepseek-r1

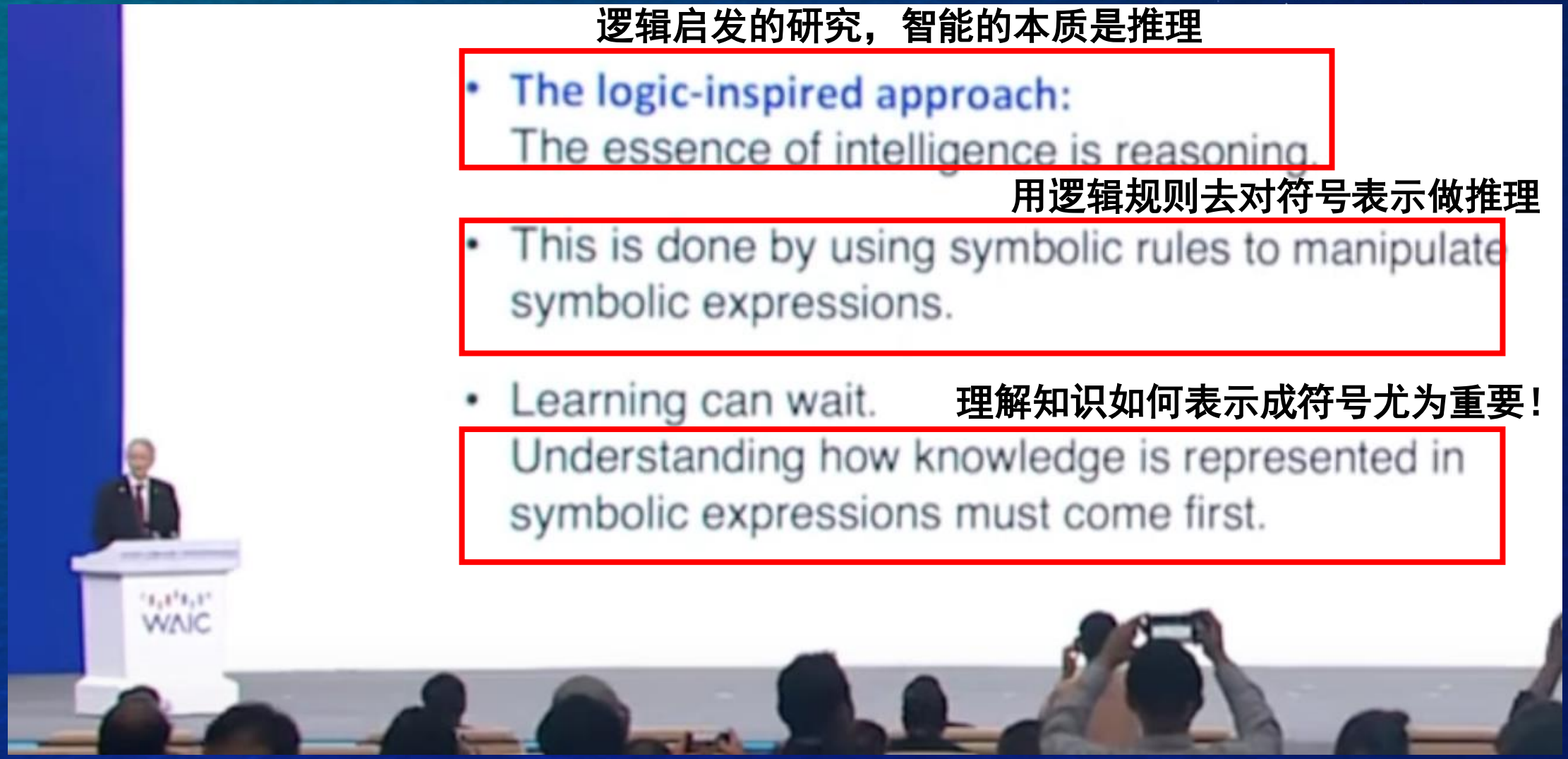
由于符号逻辑的通用性和完备性，逻辑推理能力的提升逐渐成为降低大模型幻觉的重要抓手

然而，当前的大模型逻辑推理能力不佳，限制了它们在高危+强逻辑场景下的应用

医疗咨询、法律咨询、智能工业制造等

GPT4在医学QA任务中表现30%的错误率；分析bad case可得，主要原因是忽略了药物之间的关联（冲突）或没能推理作用逻辑（酶的相互作用）

大模型逻辑表示和推理的启示（杰弗里·辛顿）



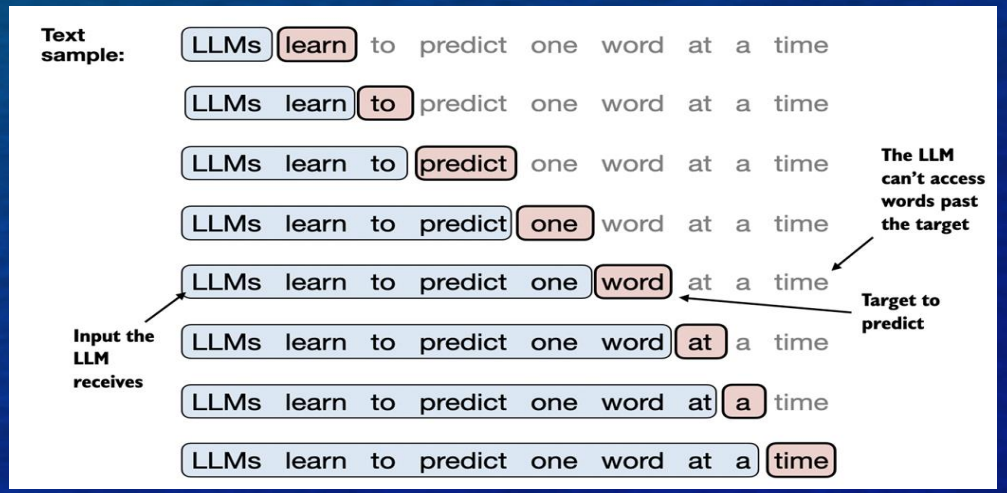
——杰弗里·辛顿（2018年图灵奖得主、2024年诺贝尔物理学奖得主），2025.7.26，世界人工智能大会，中国上海

为何大语言模型在逻辑推理中面临挑战?

大语言模型在逻辑推理方面的局限性主要源于两个关键方面：

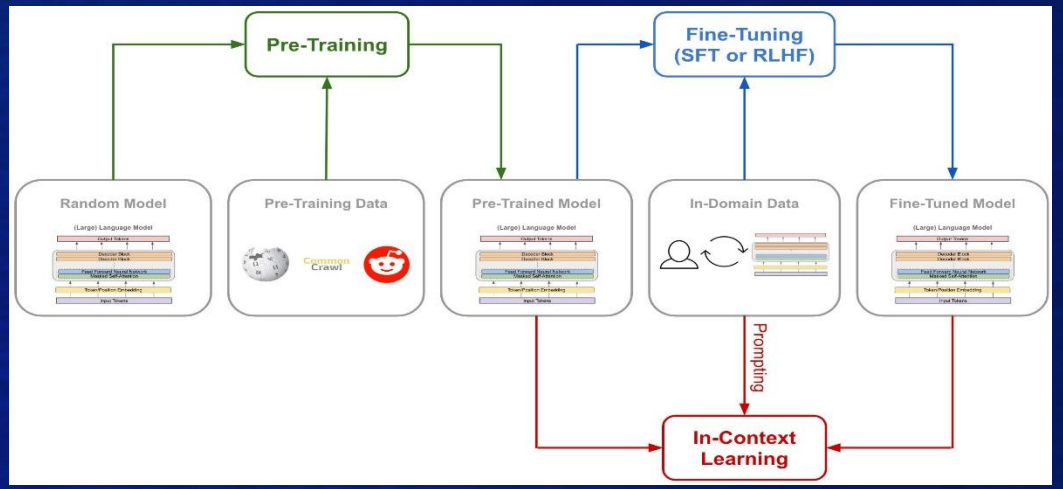
训练目标层面

- 预训练阶段的训练方法是**下一词预测**（NTP），使模型能够学习语义、语法和世界知识。
- 然而，这并**不能保证**逻辑推理能力



训练数据层面

- 预训练数据主要由**人类编写**的自然语言语料库组成。
- 训练和推理数据缺乏**高质量**的逻辑推理样本，如推导和证明



大模型逻辑推理的两大研究问题

大语言模型的逻辑推理

确保跨不同问题的**逻辑一致性**

逻辑一致性要求大语言模型在复杂推理过程中回答不同问题时不能自相矛盾。

- 问：喜鹊是鸟吗？答：是。
- 问：鸟有翅膀吗？答：是。
- 问：喜鹊有翅膀吗？答：没有。

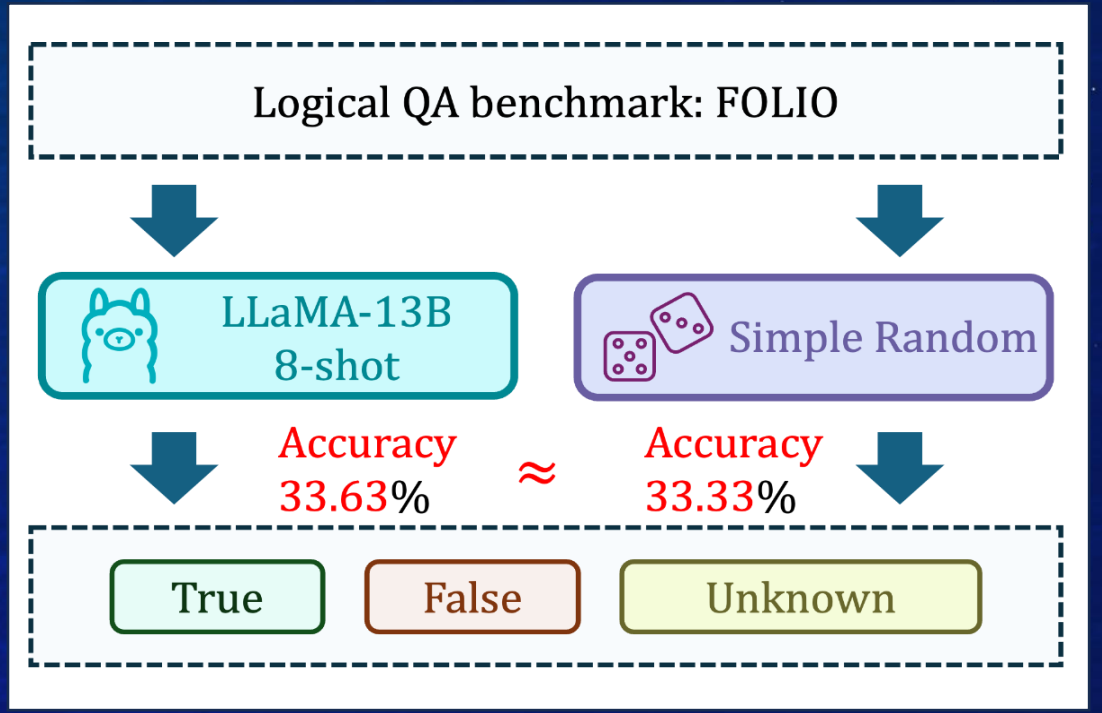
准确解答每个**逻辑问答**任务

判断某陈述是否能从给定信息中逻辑推导出来。

- **前提**：金属能导电。如果某物由铁制成，那么它就是金属。钉子由铁制成。
- $\forall x\,(Metal(x) \rightarrow Conductive(x))$，$\forall x\,(Iron(x) \rightarrow Metal(x))$，$Iron(Nails)$
- **问题**：以下陈述是真、假还是未知？
- “钉子不能导电。”
- $\neg\, Conductive(Nails)$
- **答案**：错误。

两种逻辑问答任务

- 判断一个陈述是否能**从给定信息中逻辑推导**出来（即确定其真值）。
- 从多个选项中选择**满足所有给定前提**和约束的正确选项。



大模型逻辑推理方向最新综述

发表大模型逻辑推理方向 首篇综述

Empowering LLMs with Logical Reasoning: A Comprehensive Survey

Fengxiang Cheng[1], Haoxuan Li[2,3*], Fenrong Liu[1,4], Robert van Rooij[1], Kun Zhang[3,5] and Zhouchen Lin[6,7,8*]

- 北京大学林宙辰团队，联合清华、CMU、UvA、MBZUAI 等 5 所高校，调研大模型逻辑推理的前沿工作
- 联合发布综述论文《Empowering LLMs with Logical Reasoning》，被 IJCAI 2025 Survey & Tutorial 接收
- 研究成果受邀在斯坦福大学、麻省理工大学等现场报告
- 研究成果被
  - “中国人工智能学会（CAAI）”
  - “麻省理工科技评论（MIT Technology Review）”
  - “机器之心”等媒体报道

学术分享丨北大、清华、UvA、CMU等联合发布：大模型逻辑推理能力最新综述

中国人工智能学会 2025年05月08日 12:39 北京

清华大学 Tsinghua University · 新闻 NEWS

清华主页 - 清华新闻 - 学术科研 - 正文

逻辑学研究中心联合发表大模型逻辑推理能力最新综述

清华新闻网5月13日电 当前，大模型研究正逐步从依赖扩展定律（Scaling Law）的预训练转向聚焦推理能力的后训练。鉴于符号逻辑推理的有效性与普遍性，提升大模型的逻辑推理能力成为解决幻觉问题的关键途径。

北大、清华、UvA、CMU等联合发布：大模型逻辑推理能力最新综述

机器之心 2025年05月07日 10:35 韩国 阅读 1.4万 1539

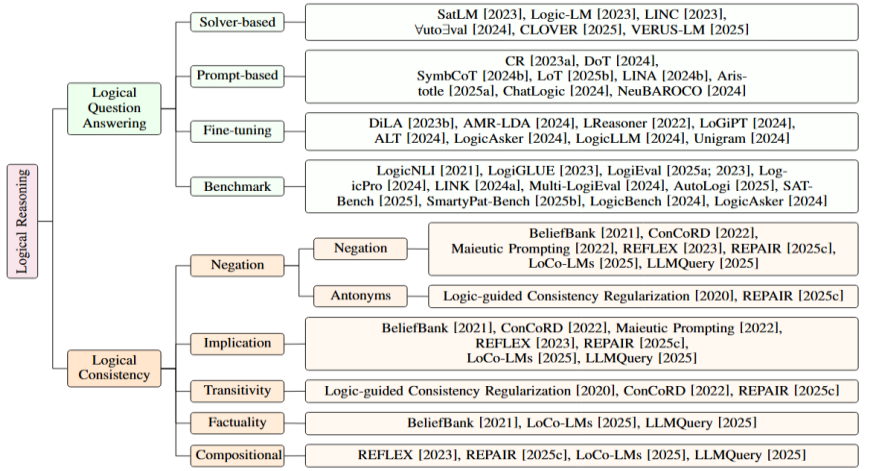


Figure 1: Our taxonomy tree of Logical Reasoning, which is primarily divided into Logical Question Answering and Logical Consistency.

综述分类图

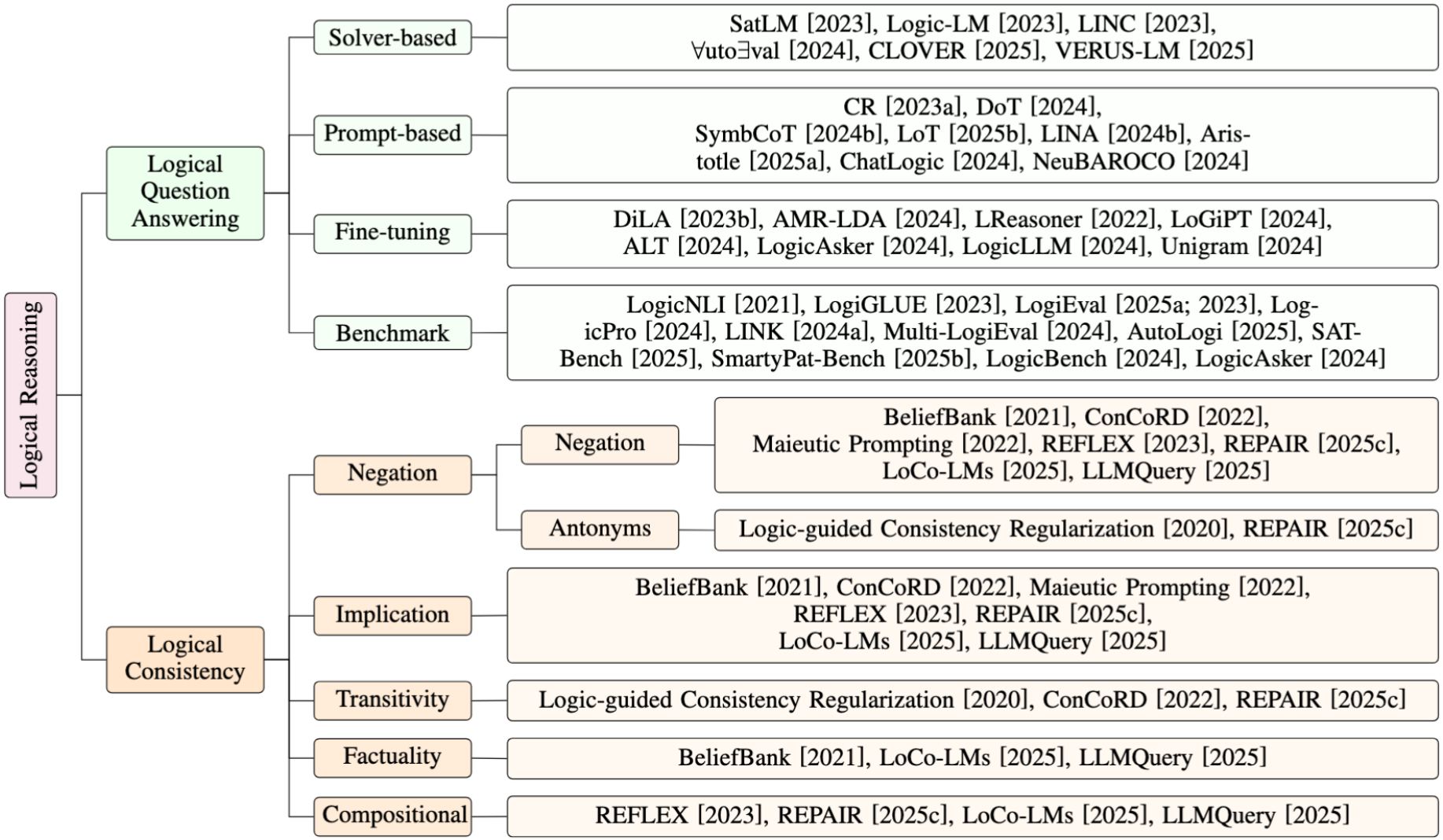


Figure 1: Our taxonomy tree of Logical Reasoning, which is primarily divided into Logical Question Answering and Logical Consistency.

步骤1：将自然语言翻译至符号语言

方法：提示大语言模型进行翻译





语言类型/风格转换较鲁棒



翻译错误

步骤2: 基于符号语言的推理

方法1: 调用逻辑求解器





最强的推理能力



翻译错误导致的较低执行率

步骤2: 基于符号语言的推理

方法2: 基于提示大模型





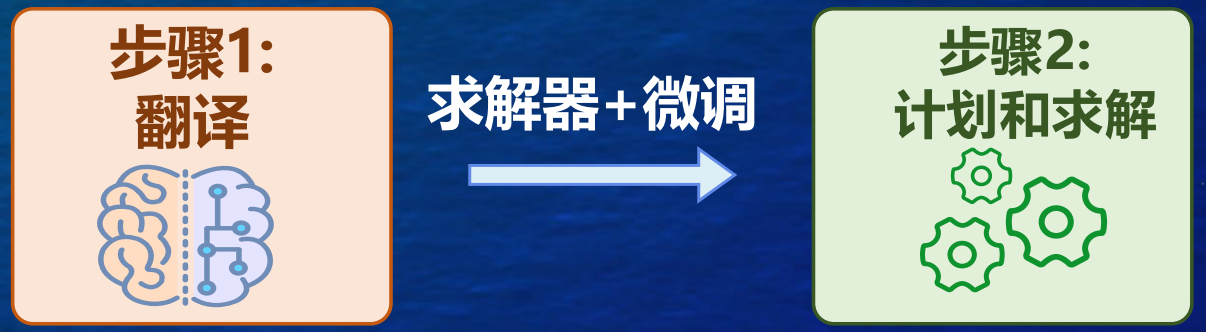
对翻译错误鲁棒



较弱的推理能力

步骤2: 基于符号语言的推理

方法3: 基于微调





直接提升大模型逻辑推理能力



依赖于高质量的合成数据和大量计算资源

大模型逻辑推理的两大研究问题

- 确保跨不同问题的**逻辑一致性(consistency)**
- 准确解答每个**逻辑问答(QA)** 任务

大语言模型的逻辑一致性

- **什么是不一致性?**
- 大语言模型逻辑一致性的基准测试
- 与自然语言推理相关的逻辑一致性
  - 面向神经模型一致性的逻辑驱动框架
- 自然语言推理赋能的问答任务
  - BeliefBank：为预训练语言模型添加记忆，系统化界定其信念
  - 通过自然语言推理增强预训练语言模型的自一致性与性能

逻辑不一致性

否定



信天翁是生物吗?



是。



信天翁不是生物吗?



是。

矛盾!

$$p \oplus \neg p, i.e., (p \vee \neg p) \wedge \neg(\neg p \wedge p)$$

逻辑不一致性

蕴涵



所有的铁 (p) 都是金属 (q) 吗?



是。

矛盾！



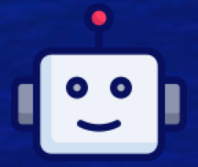
这种材料是铁 (p) 吗?



是。



这种材料是金属 (q) 吗?



不是。

$$p \to q, \quad p \nvdash q$$

逻辑不一致性

传递性



如果海啸发生，
木头会变成什么？



木头会**更**潮湿。

矛盾！



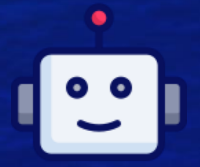
如果木头更潮湿，
风化会如何变化？



会发生**更多**的风化。



如果海啸发生，
风化会如何变化？



会发生**更少**的风化。

$$p \to q, \quad q \to r \vdash p \to r$$

逻辑不一致性

事实性



世界上最高的山是什么？



世界上最高的山是
厄尔布鲁士山。



世界上最高的山是
珠穆朗玛峰。

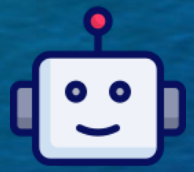
矛盾！

$$T \vdash y^*$$

逻辑不一致性

组合性


熔化是一种相变吗?


是。


冰会熔化吗?


会。


冰会经历相变吗?


会。


相变会改变质量吗?


不会。


冰的质量会改变吗?


会。

矛盾!

大语言模型的逻辑一致性

- 什么是不一致性?
- **大语言模型逻辑一致性的基准测试**
- 与自然语言推理相关的逻辑一致性
  - 面向神经网络模型一致性的逻辑驱动框架
- 自然语言推理赋能的问答任务
  - BeliefBank：为预训练语言模型添加记忆，系统化界定其信念
  - 通过自然语言推理增强预训练语言模型的自一致性与性能

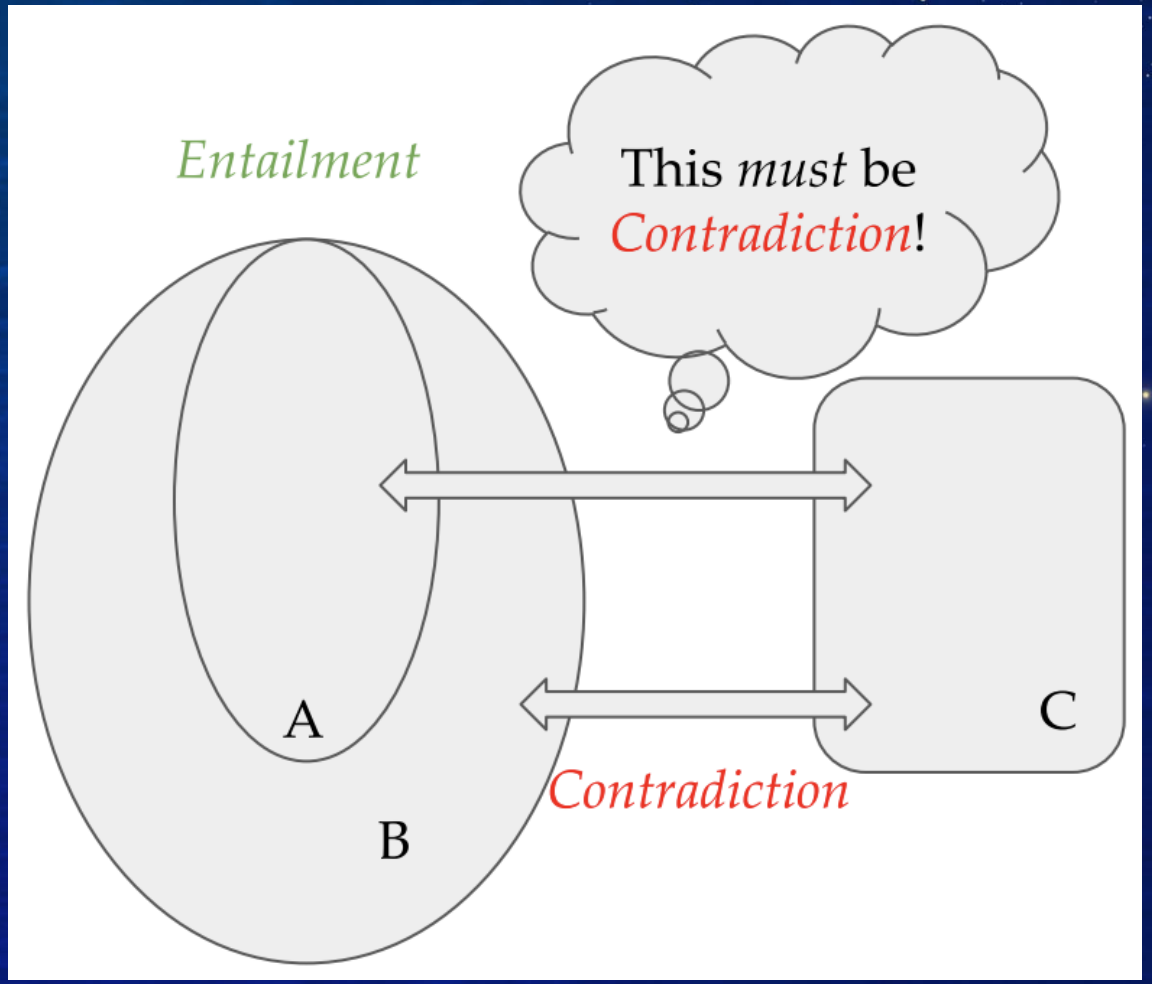
大语言模型的逻辑一致性

- 什么是不一致性?
- 大语言模型逻辑一致性的基准测试
- **与自然语言推理相关的逻辑一致性**
  - **面向神经网络模型一致性的逻辑驱动框架**
- 自然语言推理赋能的问答任务
  - BeliefBank：为预训练语言模型添加记忆，系统化界定其信念
  - 通过自然语言推理增强预训练语言模型的自一致性与性能

动机示例：人类推理

- 考虑以下句子：
  - A：Bob 在开往柏林的火车上。
  - B：Bob 正在前往柏林。
  - C：Bob 正在柏林吃饭。

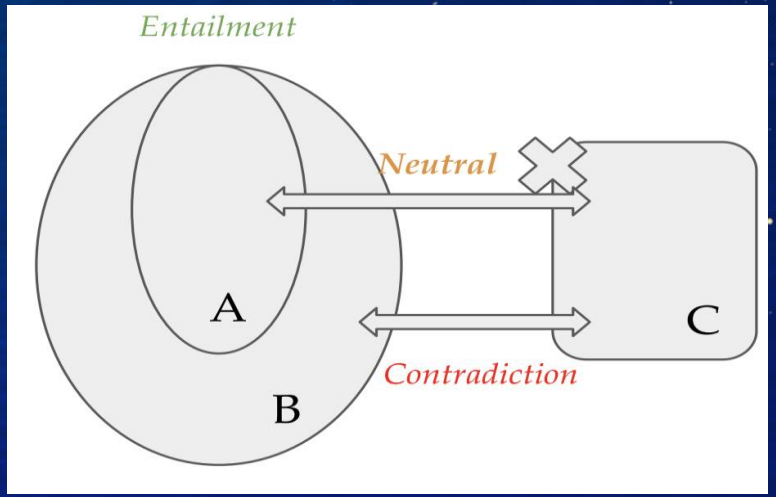
- 已知
  - (A, B)：**蕴涵（Entailment）**
  - (B, C)：**矛盾（Contradiction）**
  - 即使不看 (A, C)，我们也能推出 **(A, C) 是矛盾**



A logic-driven framework for consistency of neural models. EMNLP, 2019

动机示例：模型信念不一致的情形

- 考虑以下句子：
  - A：Bob 在开往柏林的火车上。
  - B：Bob 正在前往柏林。
  - C：Bob 正在柏林吃饭。



	Input	True	Entailment	Contradiction	Neutral
mnli_roberta-2020.06.09	(A, B)	E	**0.993**	0.000	0.006
	(B, C)	C	0.206	**0.492**	0.302
	(A, C)	C	**0.814**	0.020	0.166

A logic-driven framework for consistency of neural models. EMNLP, 2019

挑战：一致性

- 一个好的系统：既能做出正确的推理，又能保持信念的一致
- 即使准确率很高的模型，也可能是不一致的
- 应对策略？
  - 训练模型时加入这类“对抗性输入”
    - 标注成本高
    - 模型可能学不到一致性
  - 在训练中注入知识（**本文的做法**）
    - 问：能否在**后处理（post-processing）**中注入知识？

A logic-driven framework for consistency of neural models. EMNLP, 2019

挑战：一致性

- 一个好的系统：既能做出正确的推理，又能保持信念的一致
- 即使准确率很高的模型，也可能是不一致的
- 应对策略？
  - 训练模型时加入这类“对抗性输入”
    - 标注成本高
    - 模型可能学不到一致性
  - 在训练中注入知识（**本文的做法**）
    - 问：能否在**后处理（post-processing）**中注入知识？

以逻辑形式注入。

A logic-driven framework for consistency of neural models. EMNLP, 2019

方法：将知识表示为一阶逻辑

- 设x 为样本集合 D 中的一个样本（有标注或无标注）。所有约束都可统一表达为

$$\bigwedge_{(L,R)} L(x) \rightarrow R(x)$$

其中L 和R是由模型对样本x的预测所构成的布尔公式

- 前述例子中的规则可写为

$$E(A,B) \wedge C(B,C) \rightarrow C(A,C)$$

- 模型应在x_i上预测出其真实标签y_i，该规则可写为

$$\mathrm{TRUE} \rightarrow y_i(x_i)$$

这是对准确率的约束，并将与交叉熵损失（CE loss）相关联。

A logic-driven framework for consistency of neural models. EMNLP, 2019

方法：一致性规则

标注一致性: $\forall(P,H), Y \in D, \text{TRUE} \rightarrow Y(P,H)$

若数据有标注，模型就应预测出该标注

对称一致性: $\bigwedge_{(P,H)\in D} C(P,H) \leftrightarrow C(H,P)$

若 A 与 B 矛盾，则 B 也与 A 矛盾

传递一致性: $\forall(P,H,Z) \in D,$

$$\begin{aligned}
&\big(E(P,H) \wedge E(H,Z) \rightarrow E(P,Z)\big) \\
\wedge &\big(E(P,H) \wedge C(H,Z) \rightarrow C(P,Z)\big) \\
\wedge &\big(N(P,H) \wedge E(H,Z) \rightarrow \neg C(P,Z)\big) \\
\wedge &\big(N(P,H) \wedge C(H,Z) \rightarrow \neg E(P,Z)\big)
\end{aligned}$$

即前述例子

A logic-driven framework for consistency of neural models. EMNLP, 2019

方法： Softened Logic 损失

- 用预测的标签概率（如 $e(A,B)$）作为布尔决策（$E(A,B)$）的软替代
- 用三角范数（t-norm）将逻辑规则软化为损失

Name	Boolean Logic	Product	Gödel	Łukasiewicz
Negation	$\neg A$	$1-a$	$1-a$	$1-a$
T-norm	$A \wedge B$	ab	$\min(a,b)$	$\max(0,a+b-1)$
T-conorm	$A \vee B$	$a+b-ab$	$\max(a,b)$	$\min(1,a+b)$
Residuum	$A \rightarrow B$	$\min\left(1,\frac{b}{a}\right)$	$\begin{cases} 1, & \text{if b} \geq \text{a}, \\ b, & \text{else} \end{cases}$	$\min(1,1-a+b)$

A logic-driven framework for consistency of neural models. EMNLP, 2019

方法： Softened Logic 损失

标注一致性: $\forall(P,H), Y \in D, \text{TRUE} \rightarrow Y(P,H)$

$$L_{ann} = \sum_{(P,H),Y^{\star}\in D} -\log y^{\star}_{(P,H)}$$

对称一致性: $\wedge_{(P,H)\in D}\, C(P,H) \leftrightarrow C(H,P)$

$$L_{sym} = \sum_{(P,H)\in D} |\log c_{(P,H)} - \log c_{(H,P)}|$$

传递一致性:

$$\begin{aligned}
&\text{ReLU}\left(\log e(P,H) + \log e(H,Z) - \log e(P,Z)\right)\\
+&\text{ReLU}\left(\log e(P,H) + \log c(H,Z) - \log c(P,Z)\right)\\
+&\text{ReLU}\left(\log n(P,H) + \log e(H,Z) - \log\left(1 - c(P,Z)\right)\right)\\
+&\text{ReLU}\left(\log n(P,H) + \log c(H,Z) - \log\left(1 - e(P,Z)\right)\right)
\end{aligned}$$

A logic-driven framework for consistency of neural models. EMNLP, 2019

方法： Softened Logic 损失

标注一致性: $\forall(P,H), Y \in D, \text{TRUE} \rightarrow Y(P,H)$

$$L_{ann} = \sum_{(P,H),Y^\star \in D} -\log y^\star_{(P,H)}$$

对称一致性: ∧

1. 易于优化!
2. 可同时利用有标注与无标注数据!

传递一致性:

$$\begin{aligned}
&\text{ReLU}\left(\log e(P,H) + \log e(H,Z) - \log e(P,Z)\right) \\
+&\text{ReLU}\left(\log e(P,H) + \log c(H,Z) - \log c(P,Z)\right) \\
+&\text{ReLU}\left(\log n(P,H) + \log e(H,Z) - \log\left(1 - c(P,Z)\right)\right) \\
+&\text{ReLU}\left(\log n(P,H) + \log c(H,Z) - \log\left(1 - e(P,Z)\right)\right)
\end{aligned}$$

A logic-driven framework for consistency of neural models. EMNLP, 2019

不一致的度量指标

- 全局违反率 ρ
- 条件违反率 τ

# of violations

Indicator function

$$\rho = \frac{\sum_{x \in D} \left[\bigvee_{(L,R)} \neg \left(L(x) \rightarrow R(x) \right) \right]}{|D|}$$

Dataset size

$$\tau = \frac{\sum_{x \in D} \left[\bigvee_{(L,R)} \neg \left(L(x) \rightarrow R(x) \right) \right]}{\sum_{x \in D} \left[\bigvee_{(L,R)} L(x) \right]}$$

If a model tries *not* to satisfy any L, ρ can also be quite low!

$L \rightarrow R$ is a tautology if L is FALSE. Here τ *only* considers those with a true L.

A logic-driven framework for consistency of neural models. EMNLP, 2019

实验：不同训练配置下的对称性与传递性评估

Training set used

Base Model / *Fine tuned on*

Config	5% ρ_S	5% τ_S	5% ρ_T	5% τ_T	100% ρ_S	100% τ_S	100% ρ_T	100% τ_T
BERT w/ SNLI	26.3	64.4	4.9	14.8	18.6	60.3	4.7	14.9
BERT w/ MultiNLI	28.4	69.3	7.0	18.5	20.6	58.9	5.6	17.5
BERT w/ SNLI+MultiNLI	25.3	62.4	4.8	14.8	18.1	59.6	4.5	14.8
BERT w/ SNLI+MultiNLI^2	22.1	67.1	4.1	13.7	19.3	59.7	4.5	15.2
LSTM w/ SNLI+MultiNLI	25.8	69.5	9.9	21.0	16.8	53.6	5.3	16.0

If predicts (A, B) as Contradiction, then at least 60% chance it predicts (B, A) as something else.

ρ_S : *Global symmetry inconsistency*
τ_S : *Conditional symmetry inconsistency*
ρ_T : *Global transitivity inconsistency*
τ_T : *Conditional transitivity inconsistency*

A logic-driven framework for consistency of neural models. EMNLP, 2019

大语言模型的逻辑一致性

- 什么是不一致性?
- 大语言模型逻辑一致性的基准测试
- 与自然语言推理相关的逻辑一致性
  - 面向神经网络模型一致性的逻辑驱动框架
- **自然语言推理赋能的问答任务**
  - **BeliefBank：为预训练语言模型添加记忆，系统化界定其信念**
  - 通过自然语言推理增强预训练语言模型的自一致性与性能

动机：如何不微调模型，提升信念一致性

- 维持一套**一致的信念体系**是智能的关键特征，它支撑着错误调试与理性行为。

- 以往方法依赖微调，**成本高昂**，且可能破坏模型在其他任务上的**性能稳定性**。

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief. EMNLP, 2021

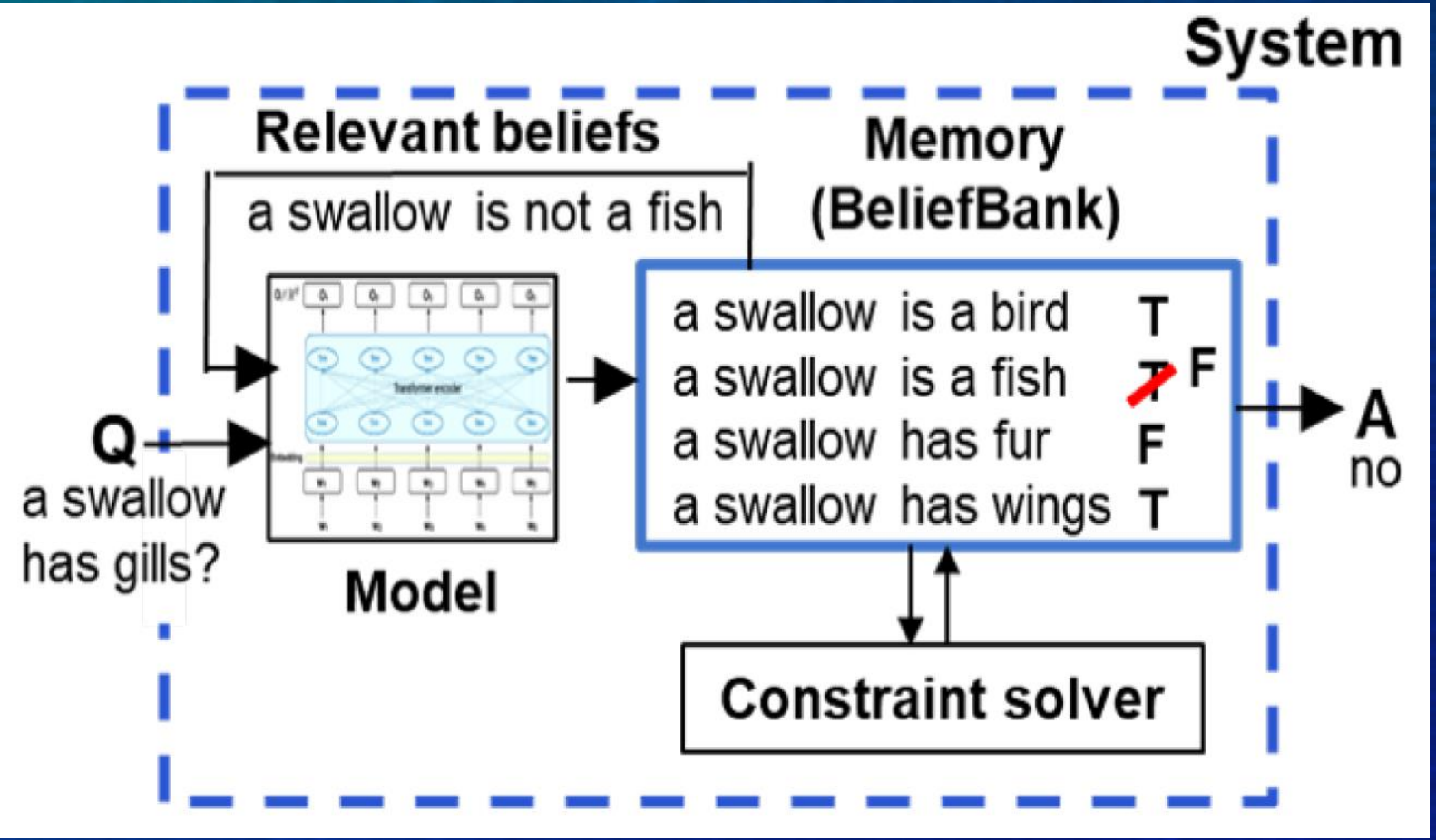
方法——总体框架

将预训练语言模型（PTLM）嵌入一个更大的系统：**该系统带有一个不断演化的、符号化的信念记忆（即信念库 BeliefBank）**，并配套两种方法来提升整个系统的信念一致性。

- 其一，**约束求解组件**（加权 MaxSAT 可满足性求解器）对信念之间的依赖关系进行推理，并修正那些与其他信念严重冲突的信念。
- 其二，**反馈组件**利用已知信念作为上下文向模型提出后续问题，使 PTLM 自身给出更准确、更一致的答案。

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief. EMNLP, 2021

总体框架：约束求解器 + 反馈机制



模型的原始回答被存入一个持久化的符号记忆（BeliefBank）中

(a) **约束求解器**翻转那些与其他信念严重冲突的信念（例如“燕子是鱼”这一信念）。

(b) **反馈机制**利用已有的相关信念（例如“燕子不是鱼”）作为查询上下文来提出新问题。

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief. EMNLP, 2021

记忆层：信念库（ BeliefBank)

▲信念库（BeliefBank）表示**整个系统关于世界的信念**（一个“心智模型”）。

▲它存储**约束**（来自 ConceptNet 与 WordNet）与**信念**（模型对一系列问题的回答）。

For example:

("a poodle is a dog", T, 0.9)

denotes the belief (strength 0.9) that "a poodle is a dog" is a true statement (T).

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief. EMNLP, 2021

约束求解：基于求解 Max SAT 问题减少信念库的知识冲突

- 给定一组信念和约束，**约束求解器**要权衡两个相互竞争的目标：
  (a) 翻转信念，使对约束的违反尽可能少；
  (b) 尽量不翻转信念，以保留模型的原始回答。

- 上述两个目标可表述为一个**加权 MaxSAT 问题**，并用 **Z3** 求解器求解。

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief. EMNLP, 2021

反馈机制：信念库相关知识作为上下文能增加准确率

- 反馈机制：向模型提问时，利用**先前相关问题**的已知答案作为辅助信息。

选择反馈信念的两种策略：

策略 1：同主题信念（On-topic）：随机选取与同一实体相关的 3 条信念。

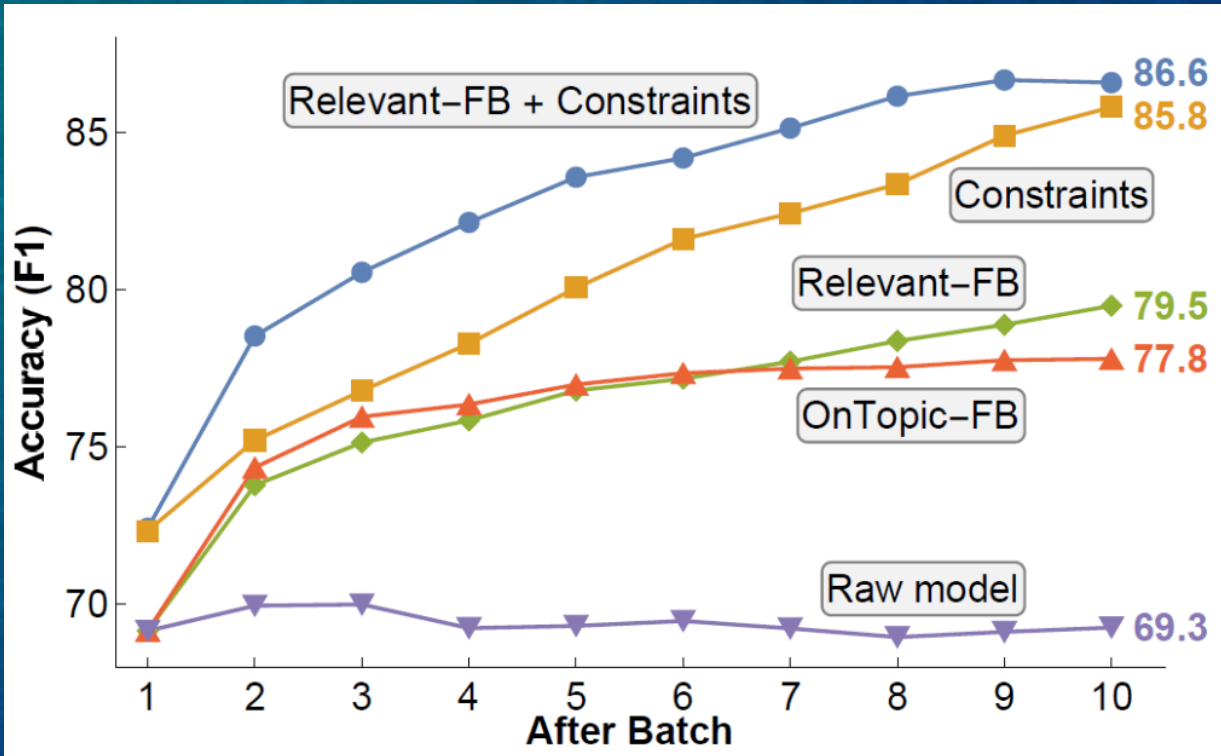
策略 2：最相关的同主题信念：利用约束图度量相关性（即最受问题 q 的可能答案影响的信念），选取冲突最强的 3 条信念作为上下文。

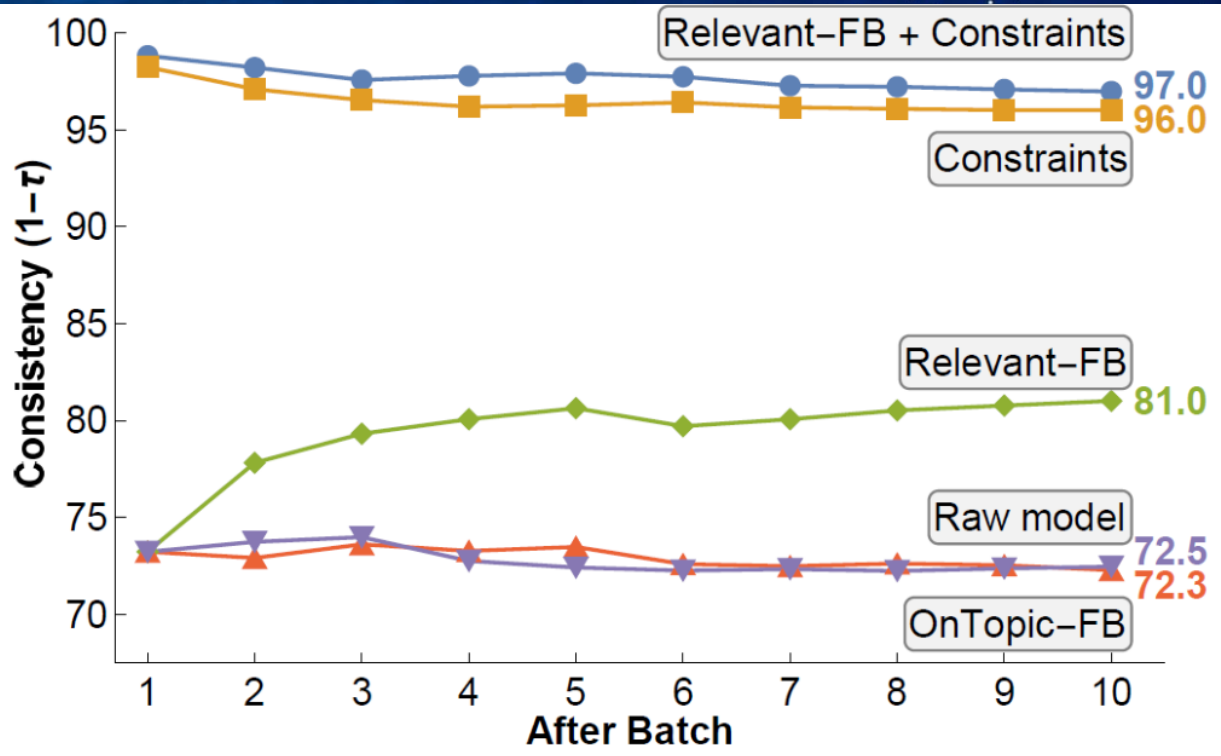
示例：

- 无反馈时：
  提问：“燕子有鳃吗？” → 模型：**有（模型回答错误）**
- 有反馈时：
  - 上下文：“燕子不是鱼。”
  - 提问：“燕子有鳃吗？” → 模型：**没有（模型回答正确）**

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief. EMNLP, 2021

实验结果：准确性与一致性同步提升

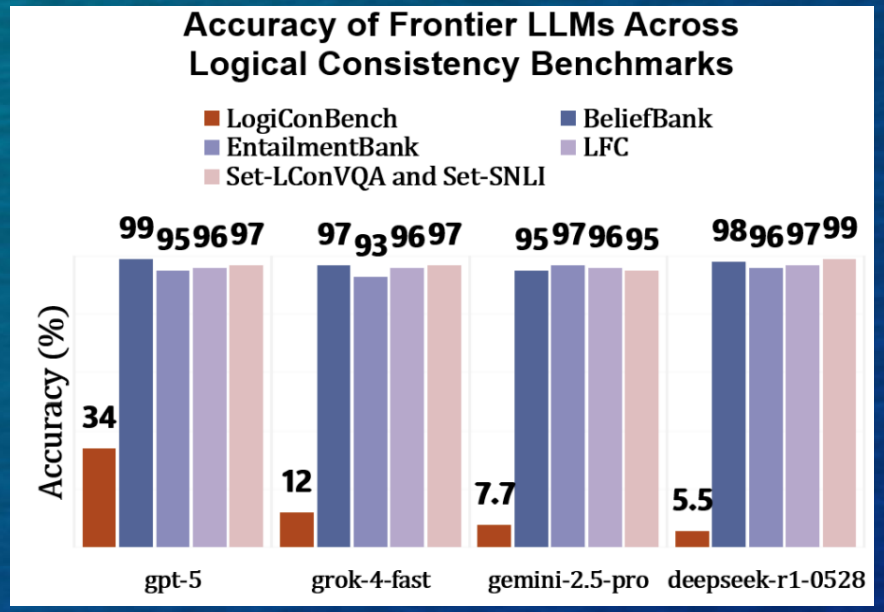




OnTopic-FB = using (randomly selected) **on-topic** feedback from old answers for new queries.
Relevant-FB = using most **relevant** on-topic feedback for new queries.
Constraints = running the constraint-solver after each batch.

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief. EMNLP, 2021

LogiConBench：每个样本由唯一逻辑规则组合程序化生成



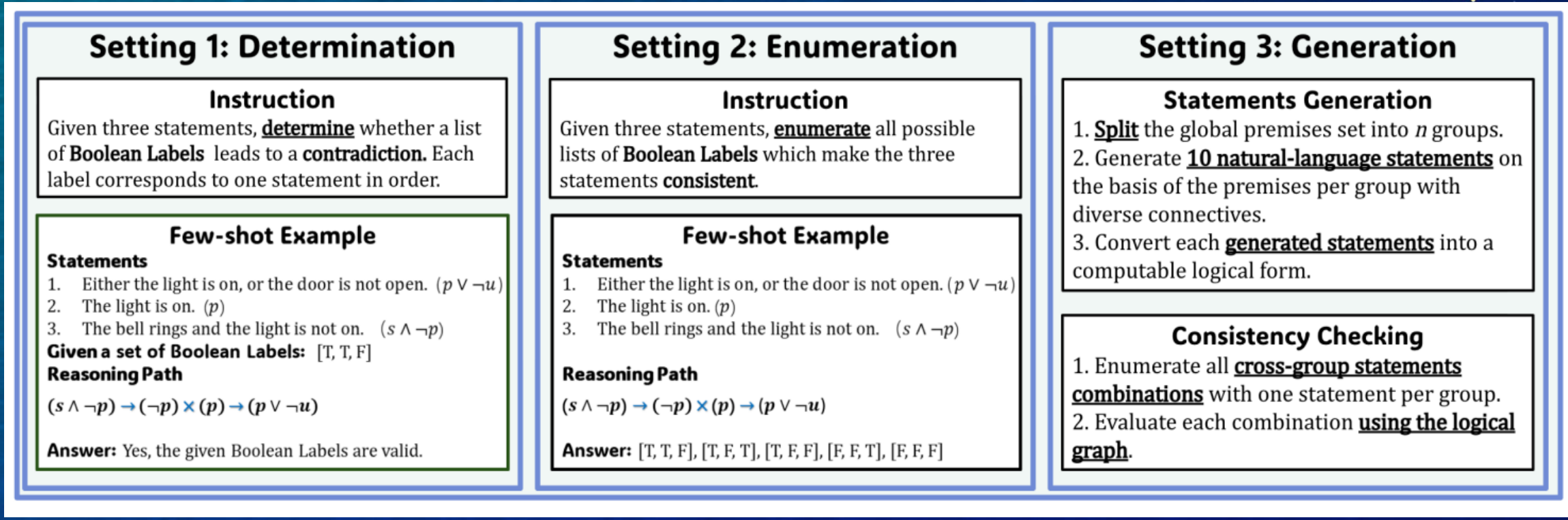
（a）前沿大语言模型在LogiConBench与现有基准准确率对比。现有基准已**趋于饱和**，而LogiConBench仍具挑战性。

（b）从规模、深度、运算符、推理路径可用性、可扩展性和规则数量方面对逻辑一致性数据集进行比较。

Dataset	Size	Depth	Operators	Reasoning Path	Scalability	# of Rule
BeliefBank	12,525	1	$\rightarrow, \neg, \leftrightarrow$	No	No	2
EntailmentBank	1,840	avg. 6	$\rightarrow$	**Yes**	No	6
LFC	~2,000	up to 4	$\wedge, \vee, \rightarrow, \neg, \leftrightarrow$	No	No	9
Set-LConVQA & Set-SNLI	13,779	up to 5	$\wedge, \vee, \rightarrow, \neg, \leftrightarrow$	No	No	51
LogiconBench	**280K**	**32**	$\wedge, \vee, \rightarrow, \neg, \leftrightarrow$	**Yes**	**Yes**	**280K**

LogiConBench: Benchmarking Logical Consistencies of LLMs. ICLR, 2026

LogiConBench测评的逻辑一致性任务（从易到难）



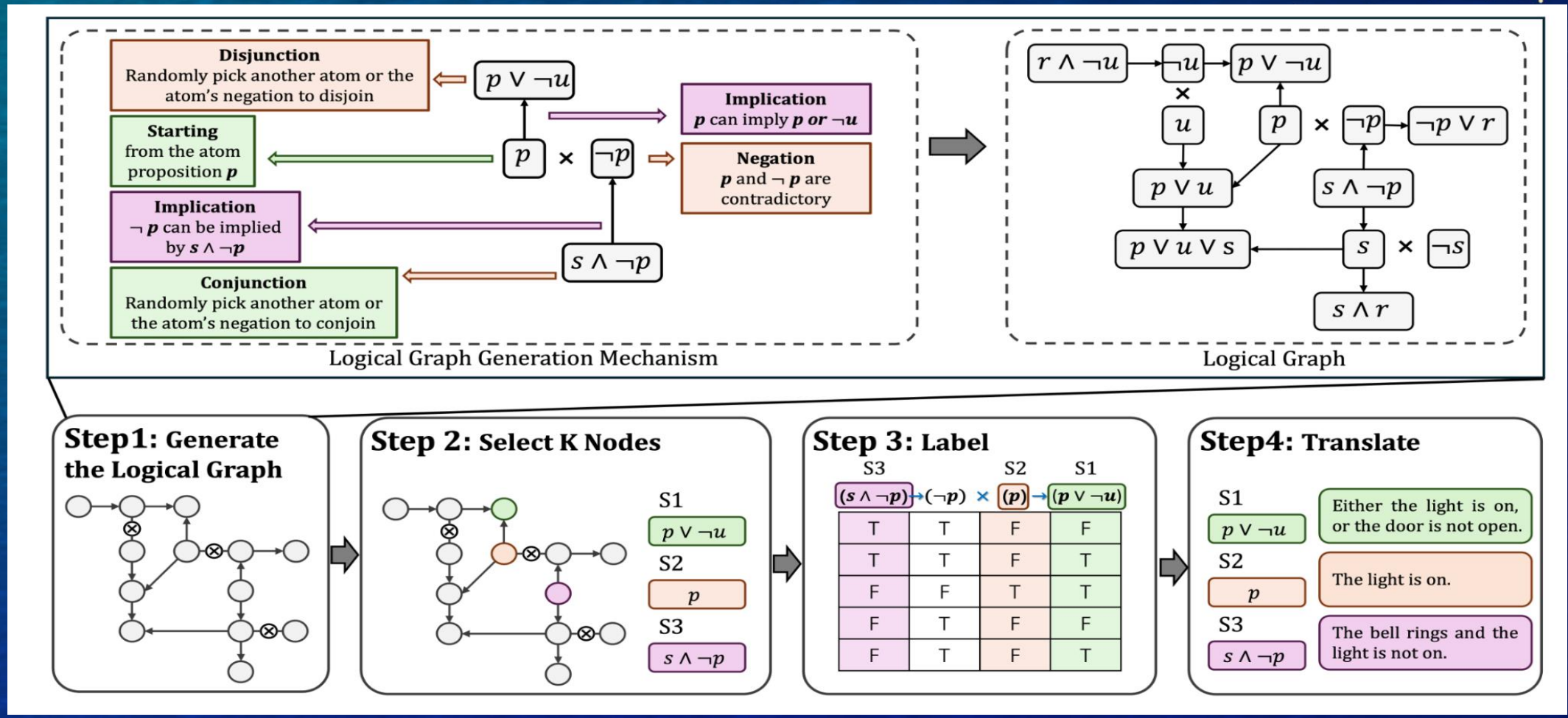
任务 1: 专注于判定分配给k个逻辑陈述的一组布尔标签是否导致矛盾，陈述规模k = 2, 3, 4, 5

任务 2: 专注于枚举任务。给定一组逻辑陈述，模型需要枚举出所有保持逻辑一致的布尔标签赋值列表

任务 3: 评估在给定n个相互一致的前提下，大语言模型是否能生成n个保持逻辑一致的新陈述

LogiConBench: Benchmarking Logical Consistencies of LLMs. ICLR, 2026

LogiConBench框架



1. 生成逻辑图　2. 选择节点　3. 标记一致集合　4. 翻译为自然语言

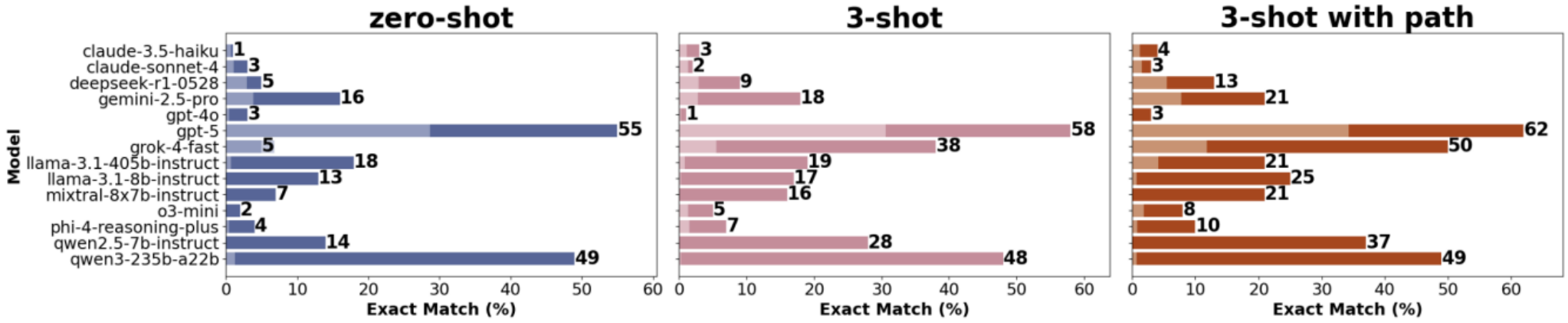
LogiConBench: Benchmarking Logical Consistencies of LLMs. ICLR, 2026

LogiConBench引入的逻辑算子和规则

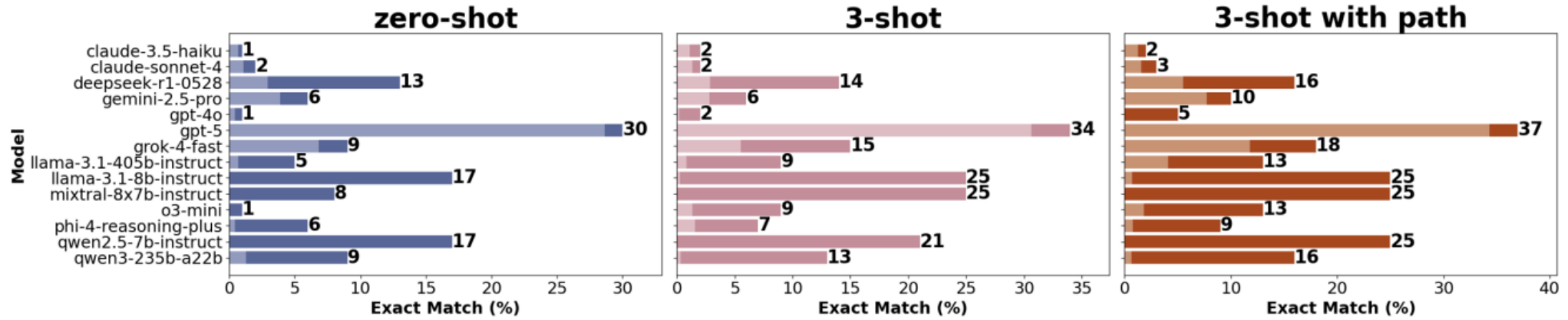
Operator	Introduction Rule	Elimination Rule
Implication ($\rightarrow$)	$I_{\rightarrow}$: If assuming φ leads (possibly through several steps) to ψ, then infer $\varphi \rightarrow \psi$ (discharge assumption, make it part of the conclusion).	$E_{\rightarrow}$ (Modus Ponens): From φ and $\varphi \rightarrow \psi$, infer ψ.
Disjunction ($\vee$)	$I_{\vee}$: From φ, infer $\varphi \vee \psi$; or from ψ, infer $\varphi \vee \psi$ (introduce $\vee$ by either side).	$E_{\vee}$: From $\varphi \vee \psi$, and the two derivations $\varphi \vdash \chi$ and $\psi \vdash \chi$, infer χ.
Conjunction ($\wedge$)	$I_{\wedge}$: From φ and ψ, infer $\varphi \wedge \psi$ (introduce $\wedge$ in the conclusion).	$E_{\wedge}$: From $\varphi \wedge \psi$, infer either φ or ψ (eliminate $\wedge$ from the premise).
Negation ($\neg$)	$I_{\neg}$: From assuming φ leads to contradiction $\perp$, infer $\neg\varphi$.	$E_{\neg}$: From φ and $\neg\varphi$, infer contradiction $\perp$.
Biconditional ($\leftrightarrow$)	$I_{\leftrightarrow}$: From $\varphi \rightarrow \psi$ and $\psi \rightarrow \varphi$, infer $\varphi \leftrightarrow \psi$.	$E_{\leftrightarrow}$: From $\varphi \leftrightarrow \psi$, infer either $\varphi \rightarrow \psi$ or $\psi \rightarrow \varphi$.

LogiConBench: Benchmarking Logical Consistencies of LLMs. ICLR, 2026

简单问题(短推理路径、短前提数量)上的逻辑一致性



(a) Task 2 on Short Reasoning Path Samples



(b) Task 2 on Short Statement Length Samples

LogiConBench在下游场景的实践价值

Benchmark	Task 1		Task 2		Task 3	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
livecodebench	0.675	0.636	0.651	0.696	0.623	0.689
infinite	0.762	0.763	0.735	0.779	0.744	0.749
aime	0.786	0.643	0.653	0.678	0.690	0.652
aa-lcr	0.624	0.598	0.614	0.695	0.603	0.665
acebench	0.632	0.640	0.526	0.688	0.714	0.702

三项任务的性能与下游基准的相关性：

- 代码生成（LiveCodeBench）
- 长上下文写作（InfiniteBench）
- 数学推理（AIME）
- 长程逻辑推理（AA-LCR）
- 智能体协作（ACE Bench）

大语言模型在真实世界下游任务上的性能与我们的基准高度相关！！

LogiConBench: Benchmarking Logical Consistencies of LLMs. ICLR, 2026

大语言模型的逻辑一致性

- 什么是不一致性?
- 大语言模型逻辑一致性的基准测试
- 与自然语言推理相关的逻辑一致性
  - 面向神经网络模型一致性的逻辑驱动框架
- **自然语言推理赋能的问答任务**
  - BeliefBank：为预训练语言模型添加记忆，系统化界定其信念
  - **通过自然语言推理增强预训练语言模型的自一致性与性能**

动机

既有方法的局限性

- **基于训练的方法：**需要**昂贵的微调**。

- **基于逻辑约束的重排序：**依赖**显式收集的逻辑关系**，这限制了可扩展性。

Enhancing Self-Consistency and Performance of Pre-Trained Language Models through Natural Language Inference. EMNLP, 2022

联合语言模型与 NLI 模型信念，解码逻辑一致的输出

基础模型生成的候选输出之间往往存在逻辑冲突。

- 候选生成：由基础预训练模型生成多个候选输出；
- NLI 约束估计：利用 NLI 模型估计候选间两两的蕴含与矛盾关系；
- 联合信念求解：联合两类模型的信念求解最一致的输出，替代贪心解码。

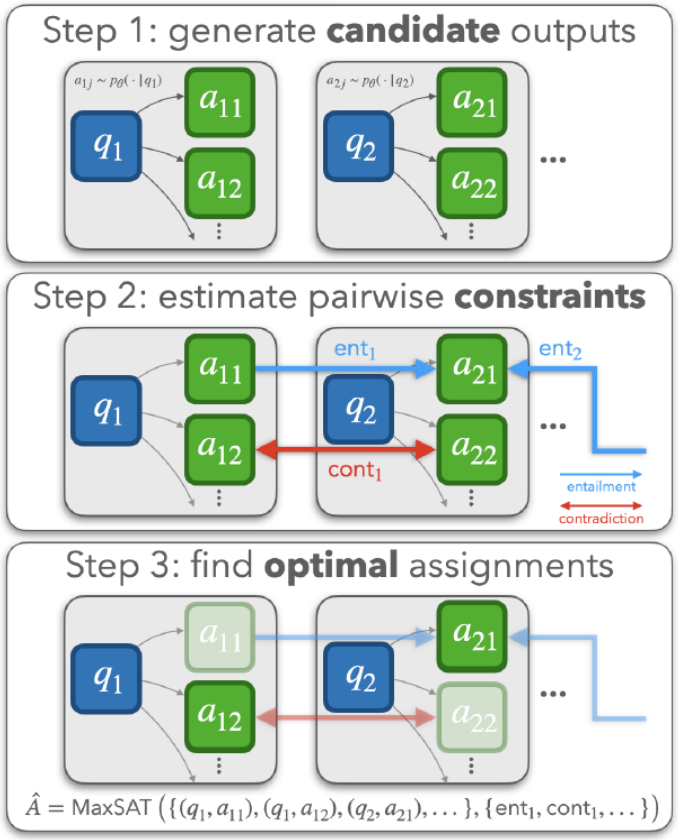
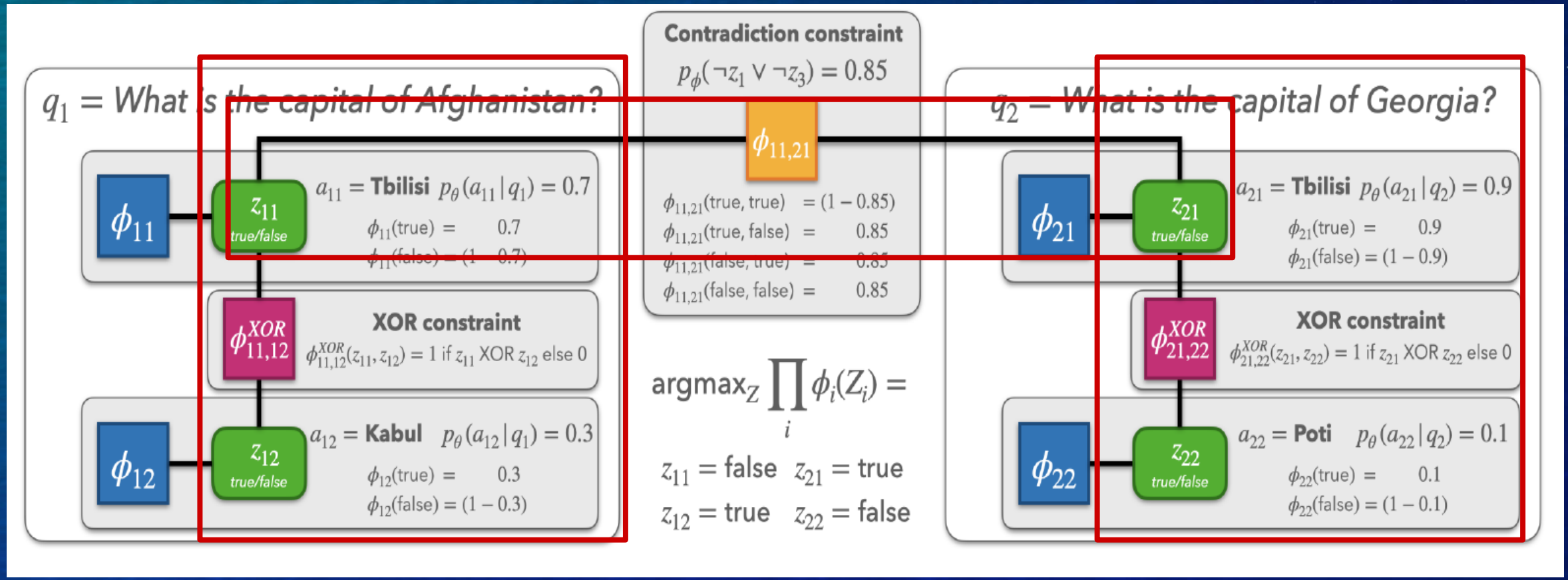


Figure 1: ConCoRD first generates candidate outputs from the base pre-trained model, then estimates soft pairwise constraints between output choices, and finally finds the most satisfactory choices of answers accounting for both the base model and NLI model's beliefs.

Enhancing Self-Consistency and Performance of Pre-Trained Language Models through Natural Language Inference. EMNLP, 2022

解码生成过程的具体示例



Enhancing Self-Consistency and Performance of Pre-Trained Language Models through Natural Language Inference. EMNLP, 2022

基于微调和基于求解器的方法的局限性

基于微调的方法

原理：用数百万事实数据点重新训练大语言模型

局限性：

1. 数据依赖性强
2. 泛化能力差

基于求解器的方法

原理：调用外部逻辑引擎进行验证

局限性：

1. 推理效率低
2. 缺乏模型优化

一种**数据依赖性低、无需外部工具、且能让大语言模型独立学习逻辑**的方法

Logically Consistent Language Models via Neuro-Symbolic Integration. ICLR, 2025

语义损失函数将逻辑约束转写为可微的优化目标

逻辑约束：

- 事实性 $z_f = z_f^*$
- 否定 $z_f \oplus z_{\tilde{f}} \Leftrightarrow (z_f \wedge \neg z_{\tilde{f}}) \vee (\neg z_f \wedge z_{\tilde{f}})$
- 蕴涵 $(z_{f_1} \rightarrow z_{f_2}) \Leftrightarrow (\neg z_{f_1} \vee z_{f_2})$

计算概率：

$$p_\theta(z_f = \top) = p_\theta(x_t = \ell_{\text{true}} \mid x_1, \ldots, x_{t-1} = \text{"信天翁是鸟吗？"})$$

惩罚不一致性：

$$\mathcal{L}(\alpha_i, p_\theta) = -\log \sum_{\mathbf{z}:\mathbf{z}\models\alpha_i} \prod_{j:\mathbf{z}\models z_{f_j}} p_\theta(z_{f_j}) \prod_{j:\mathbf{z}\models\neg z_{f_j}} \left(1 - p_\theta(z_{f_j})\right)$$

逻辑约束　　语言模型输出概率

Logically Consistent Language Models via Neuro-Symbolic Integration. ICLR, 2025

语义损失函数压低逻辑不一致组合的联合概率

以这样一条规则为例：**水仙是花** → **水仙会死**

- 大语言模型会为“水仙是花”和“水仙会死”这两个命题各赋予一个为真的概率。
- 其中只有一种逻辑不一致的结果：
  - **水仙是花** = **真**
  - **水仙会死** = **假**
- 语义损失会促使大语言模型为这一特定组合赋予非常低的概率。

Logically Consistent Language Models via Neuro-Symbolic Integration. ICLR, 2025

语义损失函数的计算示例

f_1：下雨　f_2：地湿

规则：如果下雨，则地湿

α_i: $z_{f_1} \rightarrow z_{f_2}$

假设模型输出的真值概率

$p_1 := p_\theta(z_{f_1} = T) = 0.9$

$p_2 := p_\theta(z_{f_2} = T) = 0.2$

z_{f_1}（下雨）	z_{f_2}（地湿）	组合概率	是否满足 α_i
T	T	0.9 × 0.2 = 0.18	满足
T	F	0.9 × 0.8 = 0.72	不满足
F	T	0.1 × 0.2 = 0.02	满足
F	F	0.1 × 0.8 = 0.08	满足

$$\mathcal{L}(\alpha_i, p_\theta) = -log[p_1 p_2 + (1 - p_1)p_2 + (1 - p_1)(1 - p_2)]$$

$$= -log[0.9 \times 0.2 + 0.1 \times 0.2 + 0.1 \times 0.8] = -log(0.28) \approx 1.273$$

最小化这个损失，就是最大化模型满足该逻辑约束的总概率

Logically Consistent Language Models via Neuro-Symbolic Integration. ICLR, 2025

实验结果：逻辑一致性与自一致性均得到显著提升

			CONSISTENCY			SELF-CONSISTENCY			
MODEL	TRAIN	PPL	FAC	IMP	REV	NEG	IMP	REV	AVG
LLaMa-2-7B Zero Shot		62.41	0.39	0.52	0.13	0.42	0.30	0.15	0.32
LLaMa-2-7B Few Shot		52.30	0.53	0.71	0.34	0.38	0.48	0.47	0.48
LLaMa-2-7B CoT		52.30	0.52	0.64	0.67	0.40	0.64	0.67	0.59
LLaMa-2-70B Zero Shot		44.90	0.47	0.69	0.81	0.13	0.31	0.91	0.55
LLaMa-2-7B + XENT	T1+T2	116.85	0.25	0.46	0.01	0.07	0.81	0.01	0.27
LoCo-LLaMa-2-7B (NEG)	T1	62.21	0.44	0.65	0.43	**0.96**	0.28	0.36	0.52
LoCo-LLaMa-2-7B (F-IMP)	T1	67.15	**0.99**	**0.99**	0.07	0.00	**0.99**	0.07	0.51
LoCo-LLaMa-2-7B (Super)	T1	62.23	0.74	0.77	**0.77**	0.87	0.71	**0.77**	**0.77**

LOCO-LMS 优于SFT(+XENT)

Logically Consistent Language Models via Neuro-Symbolic Integration. ICLR, 2025

量化与提升偏好中的逻辑一致性

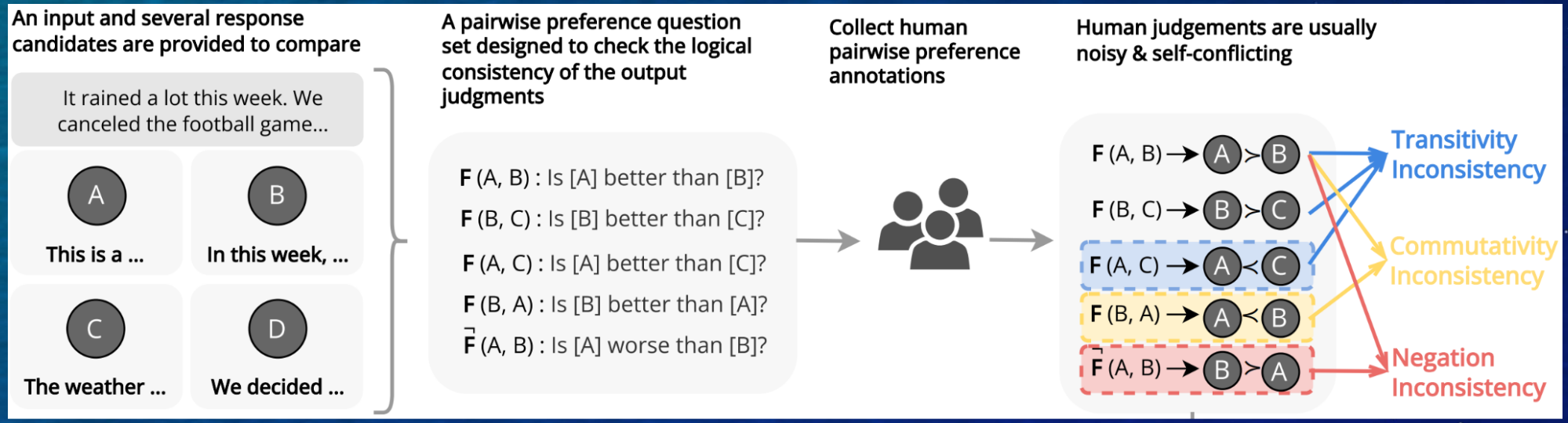
- 大语言模型在**可靠性**与**可信赖性**方面仍存在关键挑战
  - 一个可靠的系统，其**预测应当是一致的**

    -> 从而带来可解释、可处理的决策

- 如何**定义**与**量化**逻辑偏好一致性?当前的大语言模型有多一致？如何提升一致性？其影响是什么？

Aligning with Logic Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models. ICML, 2025 Spotlight

逻辑一致性究竟是什么?



- 逻辑一致性有三个关键方面:
  传递性、交换性、否定不变性。
- 对于一个项的集合 $X = \{x_1, x_2, \ldots, x_N\}$. 可以用大语言模型预测项之间某类有向关系, 例如:
  - **响应候选**之间的**偏好关系**
  - **事件**之间的**时序或因果次序关系**

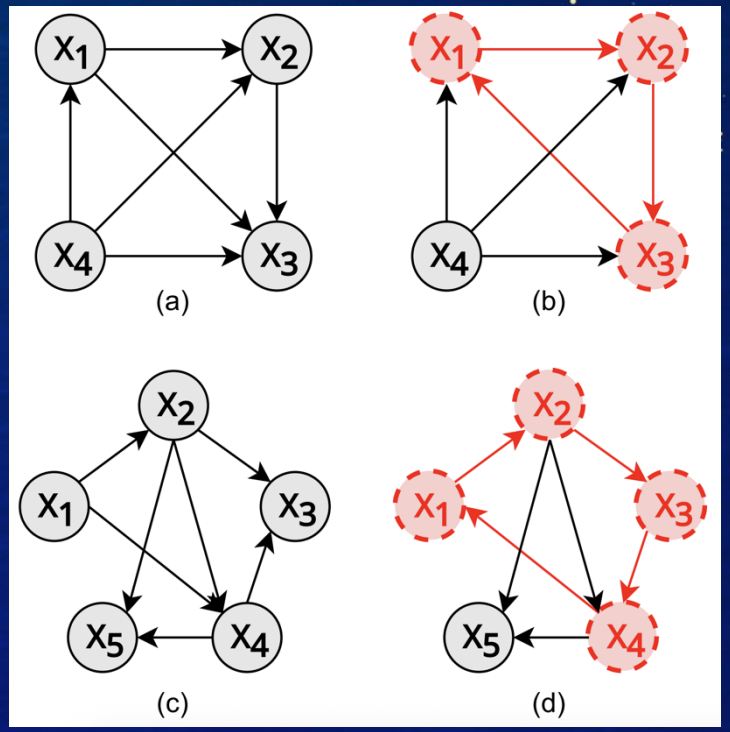
Aligning with Logic Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models. ICML, 2025 Spotlight

量化逻辑不一致性：传递性

- **传递性**是指：当系统相信 A > B 且 B > C 时，它也应预测A > C。
- 各个**项**及其**偏好关系**构成关系图上的**节点与边**。
- 非传递关系对应图中的环。
- 因此，传递性被量化为：

$$s_{tran}(K) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}_{acyclic}\left(S_i^K\right),$$

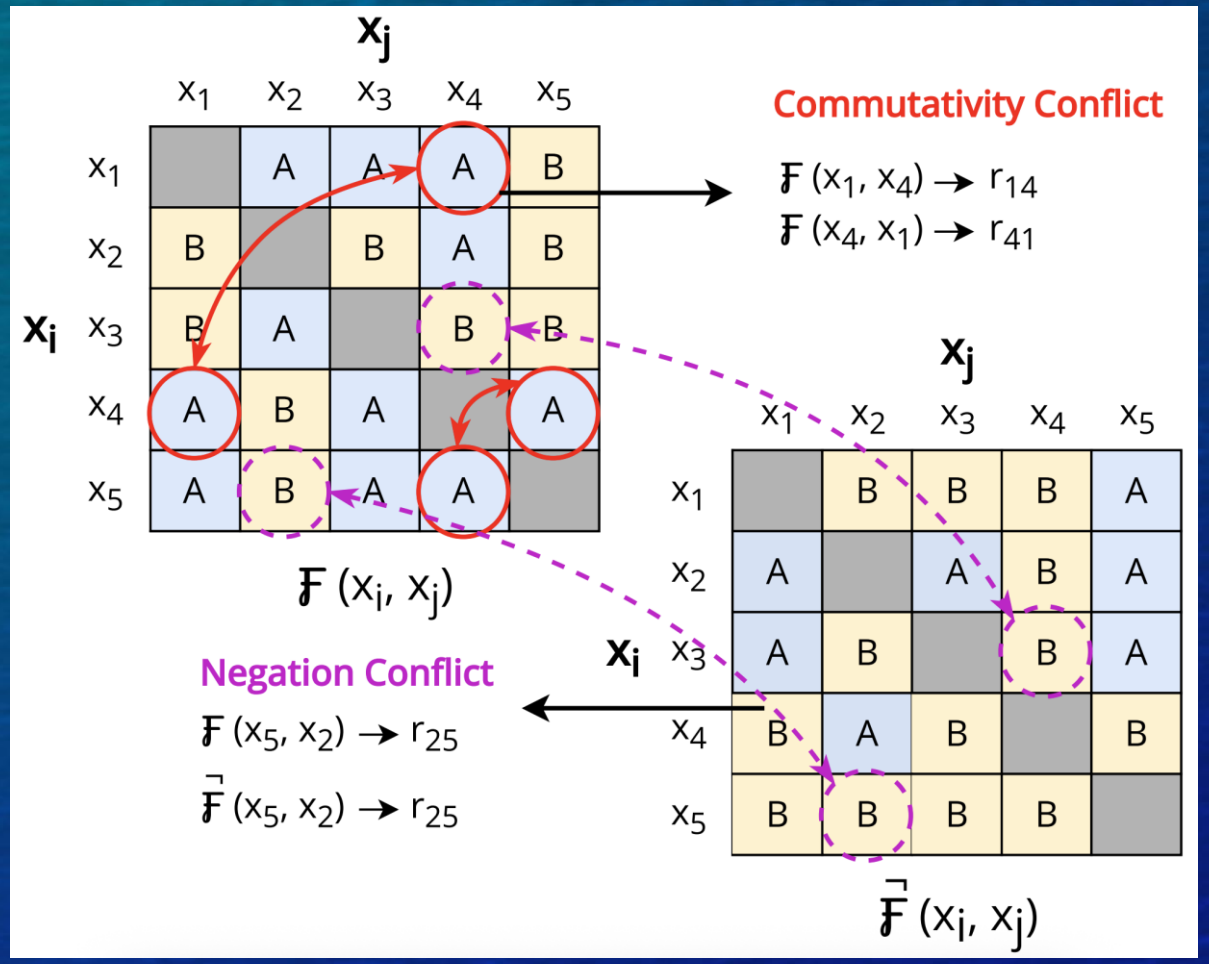
- 其中 S_i^K 是大小为 K 的子图,若该子图不含环则 $\mathbb{I}_{acyclic} = 1$ 。



检测图中的环很简单！

Aligning with Logic Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models. ICML, 2025 Spotlight

量化交换性与否定不变性



- **交换性**（也称对称性）是指以相反次序比较各项时的一致性，

  例如, ($x_1$vs x_4) and (x_4 vs x_1)

- **否定不变性**是指使用带否定关系表述的提示比较各项时的一致性，

  例如, ($x_5 > x_2$) and ($x_2 < x_5$)

Aligning with Logic Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models. ICML, 2025 Spotlight

实验结果：不同模型的逻辑不一致性被精确量化

Table 1: Logical consistency evaluation results. We report human accuracy (H.), transitivity over 5 items ($s_{tran}(5)$), commutativity (s_{comm}) and negation invariance (s_{neg}), all measured in accuracy.

Models	SummEval (Coh)				NovelEval				CaTeRS			
	H.	$s_{tran}(5)$	s_{comm}	s_{neg}	H.	$s_{tran}(5)$	s_{comm}	s_{neg}	H.	$s_{tran}(5)$	s_{comm}	s_{neg}
Direct Judgements												
Llama-2-7B	57.5	88.3	57.5	66.2	57.5	68.1	57.5	78.1	61.9	88.4	56.0	49.7
Llama-2-13B	58.3	86.6	59.3	84.0	58.3	88.2	62.9	76.9	65.6	95.3	70.8	54.5
Llama-3-8B	67.8	91.0	76.1	48.9	60.6	73.0	73.3	79.1	73.1	88.2	79.9	63.3
Mistral-7B	63.6	95.1	59.9	51.2	60.1	90.5	68.0	82.1	70.2	95.9	73.9	76.9
Zephyr-7B-beta	61.3	87.8	52.8	74.1	60.5	91.5	63.8	86.5	70.6	93.5	77.8	80.8
Phi-3-mini	65.6	92.8	66.9	75.1	59.7	92.5	55.5	39.4	60.2	**97.2**	85.5	73.3
Phi-3-medium	68.8	**96.2**	71.0	78.1	62.7	93.7	76.3	77.6	67.2	93.4	87.9	84.1
Gemma-2-9B	**73.8**	94.8	**78.1**	78.2	63.6	89.9	85.3	88.2	72.4	96.0	86.5	66.9
GPT-3.5-0125	66.3	82.5	67.5	65.8	61.2	89.8	71.6	83.4	69.3	86.2	66.7	81.7
Deepseek-chat-v2	72.7	93.1	72.8	**84.9**	**65.7**	**95.1**	**86.7**	**89.2**	**73.5**	93.8	**91.9**	**91.1**
CoT Prompting												
Llama-2-7B	55.0	42.6	45.7	68.9	56.5	50.8	66.7	60.1	59.5	40.5	56.7	56.1
Llama-2-13B	66.6	55.8	63.2	40.3	59.2	60.8	71.0	56.4	60.6	61.1	63.5	74.3
Llama-3-8B	67.2	65.0	64.8	61.2	60.2	69.3	76.8	72.4	67.9	62.2	46.1	78.6
Mistral-7B	61.6	64.7	65.3	58.0	58.5	64.9	73.2	77.7	60.3	66.1	69.0	79.1
Zephyr-7B-beta	58.4	46.4	51.7	74.0	60.0	64.5	70.6	71.2	65.7	69.3	78.4	77.3
Phi-3-mini	65.1	67.9	61.9	43.0	56.5	61.2	47.8	80.5	58.4	81.2	82.6	78,7
Phi-3-medium	69.1	**82.3**	72.5	58.9	**63.1**	89.0	87.3	79.3	54.5	81.4	87.4	85.9
Gemma-2-9B	73.5	74.0	71.9	81.3	62.1	89.9	85.9	82.3	63.5	**83.8**	87.1	69.9
GPT-3.5-0125	63.9	63.3	64.8	68.4	59.3	79.3	73.4	67.5	62.8	68.6	70.5	74.9
Deepseek-chat-v2	**74.9**	79.4	**76.1**	**82.8**	61.5	**94.2**	**90.4**	**83.0**	**69.1**	81.6	**88.2**	**88.4**

Aligning with Logic Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models.
ICML, 2025 Spotlight

实验结果：不同模型的逻辑不一致性被精确量化

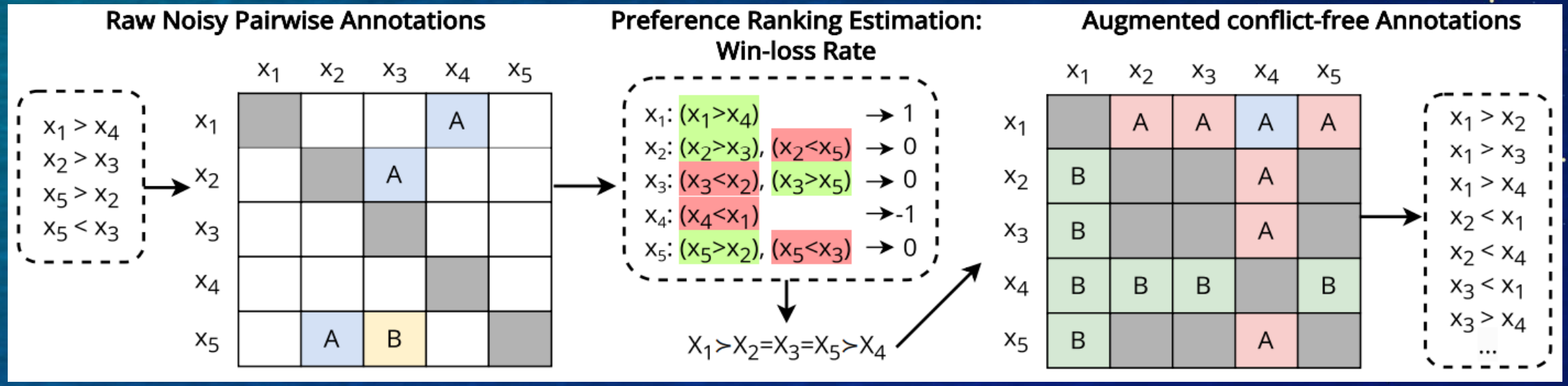
Table 1: Logical consistency evaluation results. We report human accuracy (H.), transitivity over 5 items ($s_{tran}(5)$), commutativity (s_{comm}) and negation invariance (s_{neg}), all measured in accuracy.

思维链（CoT）通常不能提升一致性

M												
											s_{comm}	s_{neg}
Direct												
Llama-2-7B	57.5	88.3	57.5	66.2	57.5	68.1	57.5	78.1	61.9	88.4	56.0	49.7
Llama-2-13B	58.3	86.6	59.3	84.0	58.3	88.2	62.9	76.9	65.6	95.3	70.8	54.5
Llama-3-8B	67.8	91.0	76.1	48.9	60.6	73.0	73.3	79.1	73.1	88.2	79.9	63.3
Mistral-7B	63.6	95.1	59.9	51.2	60.1	90.5	68.0	82.1	70.2	95.9	73.9	76.9
Zephyr-7B-beta	61.3	87.8	52.8	74.1	60.5	91.5	63.8	86.5	70.6	93.5	77.8	80.8
Phi-3-mini	65.6	92.8	66.9	75.1	59.7	92.5	55.5	39.4	60.2	**97.2**	85.5	73.3
Phi-3-medium	68.8	**96.2**	71.0	78.1	62.7	93.7	76.3	77.6	67.2	93.4	87.9	84.1
Gemma-2-9B	**73.8**	94.8	**78.1**	78.2	63.6	89.9	85.3	88.2	72.4	96.0	86.5	66.9
GPT-3.5-0125	66.3	82.5	67.5	65.8	61.2	89.8	71.6	83.4	69.3	86.2	66.7	81.7
Deepseek-chat-v2	72.7	93.1	72.8	**84.9**	**65.7**	**95.1**	**86.7**	**89.2**	**73.5**	93.8	**91.9**	**91.1**
CoT Prompting												
Llama-2-7B	55.0	42.6	45.7	68.9	56.5	50.8	66.7	60.1	59.5	40.5	56.7	56.1
Llama-2-13B	66.6	55.8	63.2	40.3	59.2	60.8	71.0	56.4	60.6	61.1	63.5	74.3
Llama-3-8B	67.2	65.0	64.8	61.2	60.2	69.3	76.8	72.4	67.9	62.2	46.1	78.6
Mistral-7B	61.6	64.7	65.3	58.0	58.5	64.9	73.2	77.7	60.3	66.1	69.0	79.1
Zephyr-7B-beta	58.4	46.4	51.7	74.0	60.0	64.5	70.6	71.2	65.7	69.3	78.4	77.3
Phi-3-mini	65.1	67.9	61.9	43.0	56.5	61.2	47.8	80.5	58.4	81.2	82.6	78,7
Phi-3-medium	69.1	**82.3**	72.5	58.9	**63.1**	89.0	87.3	79.3	54.5	81.4	87.4	85.9
Gemma-2-9B	73.5	74.0	71.9	81.3	62.1	89.9	85.9	82.3	63.5	**83.8**	87.1	69.9
GPT-3.5-0125	63.9	63.3	64.8	68.4	59.3	79.3	73.4	67.5	62.8	68.6	70.5	74.9
Deepseek-chat-v2	**74.9**	79.4	**76.1**	**82.8**	61.5	**94.2**	**90.4**	**83.0**	**69.1**	81.6	**88.2**	**88.4**

Aligning with Logic Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models. ICML, 2025 Spotlight

使用精炼与增强的偏好数据指导指令微调



在指令微调过程中提升逻辑一致性的样本数据精炼与增强框架。

- 原始的人类偏好数据通常带有噪声，且存在**自我冲突**。
- 首先，利用胜-负率等排序聚合方法，从带噪标注中估计出一个偏序；
- 随后，从该偏序中拆分出所有可用的成对比较，用于指令微调。

Aligning with Logic Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models. ICML, 2025 Spotlight

实验结果：传递性、交换性与否定性均显著提升

Table 2: Statistics at each augmentation stage. # Data means data size and AvgComp/Inst for average comparisons per instance. REPAIR-ed data represents data augmented by our REPAIR framework and Neg. denotes further augmentation with negated relation comparisons.

Data	$s_{tran}(5)$	s_{comm}	# Data	AvgComp/Inst
Raw data	98.4	-	14.8K	6.29
Perturbed data	87.6	-	14.8K	6.29
REPAIR-ed data	1	1	30.9K	13.2
REPAIR-ed data+Neg.	1	1	61.8K	26.4

Table 3: Instruction-tuning of Llama-3 Instruct (8B) using different variants (see Section 4.2). Only training on negated relations can enhance negation invariance.

Models	Summarize from Feedback			
	H.	$s_{tran}(5)$	s_{comm}	s_{neg}
Zero-shot inference	64.3	81.1	70.2	63.8
Raw data	74.4	93.1	89.0	60.6
Perturbed data	70.3	91.9	88.4	61.0
REPAIR-ed data	70.1	95.4	91.2	60.8
REPAIR-ed data + Neg.	69.7	95.2	90.5	87.9

- 性能
  - 使用精炼与增强后的数据集进行训练显著**提升传递性与交换性**，同时保持人类对齐水平
  - 只有在否定关系上进一步训练才能**提升否定不变性** (反转诅咒!)

Aligning with Logic Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models. ICML, 2025 Spotlight

大模型逻辑推理的两大研究问题

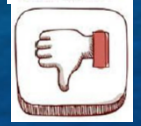
- 确保跨不同问题的**逻辑一致性(consistency)**
- 准确解答每个**逻辑问答(QA)** 任务

多智能体协同校验缓解符号翻译瓶颈

基于指令的方法流程

引导提示 → 推理+翻译

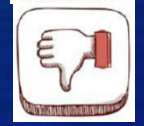

翻译过程中无信息丢失


推理存在幻觉现象

基于求解器的方法流程

步骤1：翻译 → 求解器 → 步骤2：规划与求解


强大的推理能力


因翻译错误导致执行率低

本方法动机

为什么大模型神经符号推理需要多智能体?

- **翻译阶段：**智能体在自然语言（NL）到符号语言（SL）的翻译过程中进行交叉核验与协同编辑，拓宽信息瓶颈→提升翻译质量。
- **求解阶段：**智能体结合答案+推理过程（求解器证明）进行辩论，随后通过多数投票→提高准确性与稳定性。

MAD-Logic: Multi-Agent Debate Enhances Symbolic Translation and Reasoning. ICLR, 2026

符号表示：不同逻辑语言适配不同推理任务

Problem	Formulation	Example		Solver	Dataset
		NL Sentence	Symbolic Formulation		
Deductive Reasoning	LP	If the circuit is complete and the circuit has the light bulb then the light bulb is glowing.	Complete(Circuit, True)∧ Has(Circuit, LightBulb) → Glowing(LightBulb, True)	Pyke	ProntoQA, ProofWriter
First-Order Logic	FOL	A Czech person wrote a book in 1946.	$\exists x_2 \exists x_1(\texttt{Czech}(x_1) \wedge \texttt{Author}(x_2, x_1) \wedge \texttt{Book}(x_2) \wedge \texttt{Publish}(x_2, 1946))$	Prover9	FOLIO
Constraint Satisfaction	CSP	On a shelf, there are five books. The blue book is to the right of the yellow book.	blue_book $\in \{1,2,3,4,5\}$ yellow_book $\in \{1,2,3,4,5\}$ blue_book $>$ yellow_book	python-constraint	LogicalDeduction
Analytical Reasoning	SAT	Xena and exactly three other technicians repair radios	repairs(Xena, radios) ∧ Count([t:technicians], t $\neq$ Xena ∧ repairs(t, radios))) == 3)	Z3	AR-LSAT

MAD-Logic: Multi-Agent Debate Enhances Symbolic Translation and Reasoning. ICLR, 2026

多智能体辩论兼顾符号严谨性与语言鲁棒性

翻译阶段

先前的工作：将自然语言（NL）问题翻译成特定的符号语言（SL）（例如LP、FOL、SAT等）。
然而，每种符号语言都有其自身的优缺点。因此，我们采用多智能体辩论的方式。

动机及贡献：不同的符号语言能够捕捉到原始自然语言的不同重要特征。
如果将同一自然语言翻译成多种符号语言，可以相互改进来提升彼此的翻译效果。

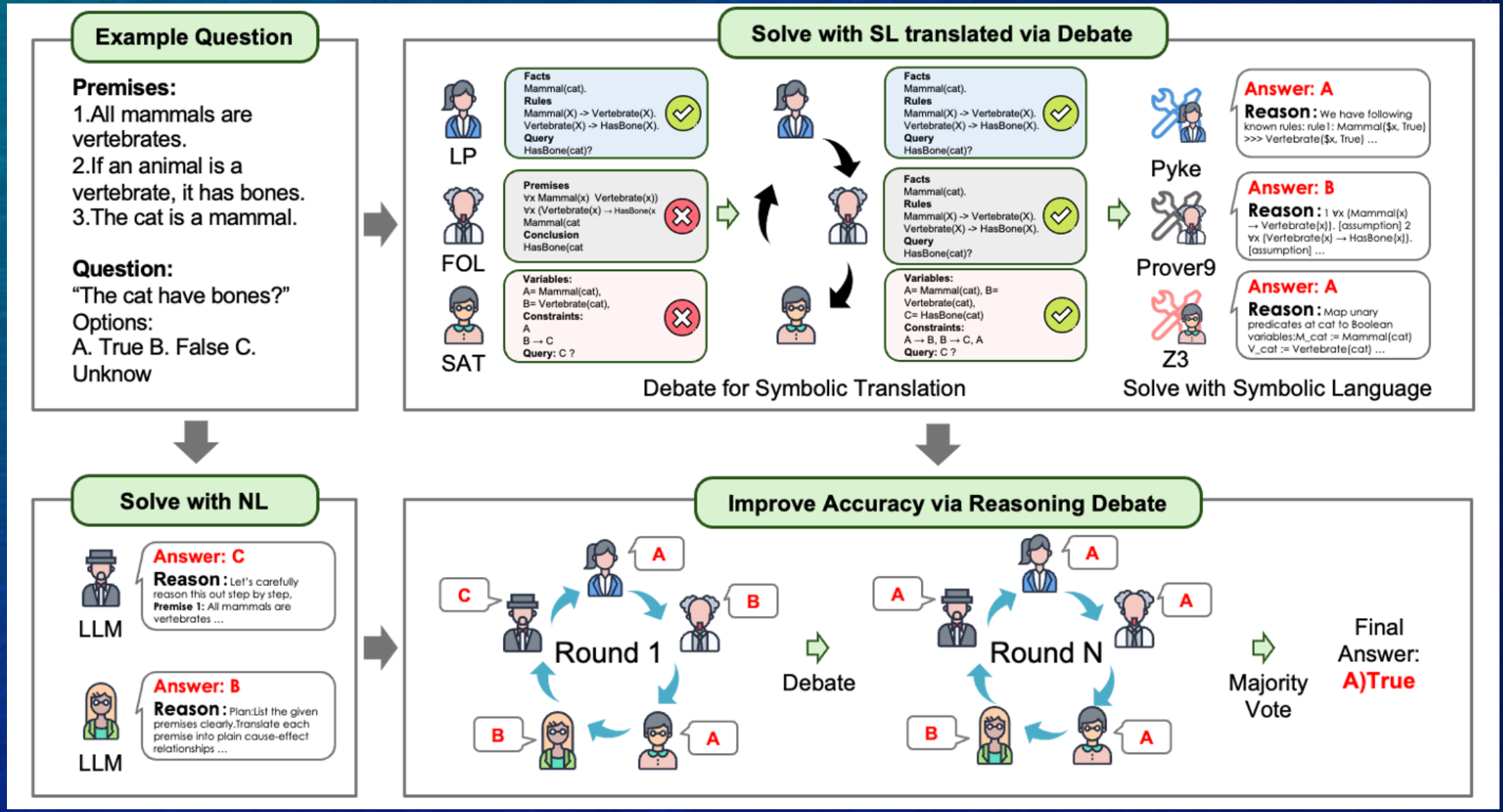
推理阶段

先前的工作：要么通过 SL 求解器进行神经符号推理，要么提示 LLM 进行神经符号推理

动机及贡献：SL求解器--推理能力强，鲁棒性弱。LLM提示--鲁棒性强，推理能力弱。
因此，我们引入多智能体辩论来综合两种方法的优势，从而在推理阶段实现最佳性能。

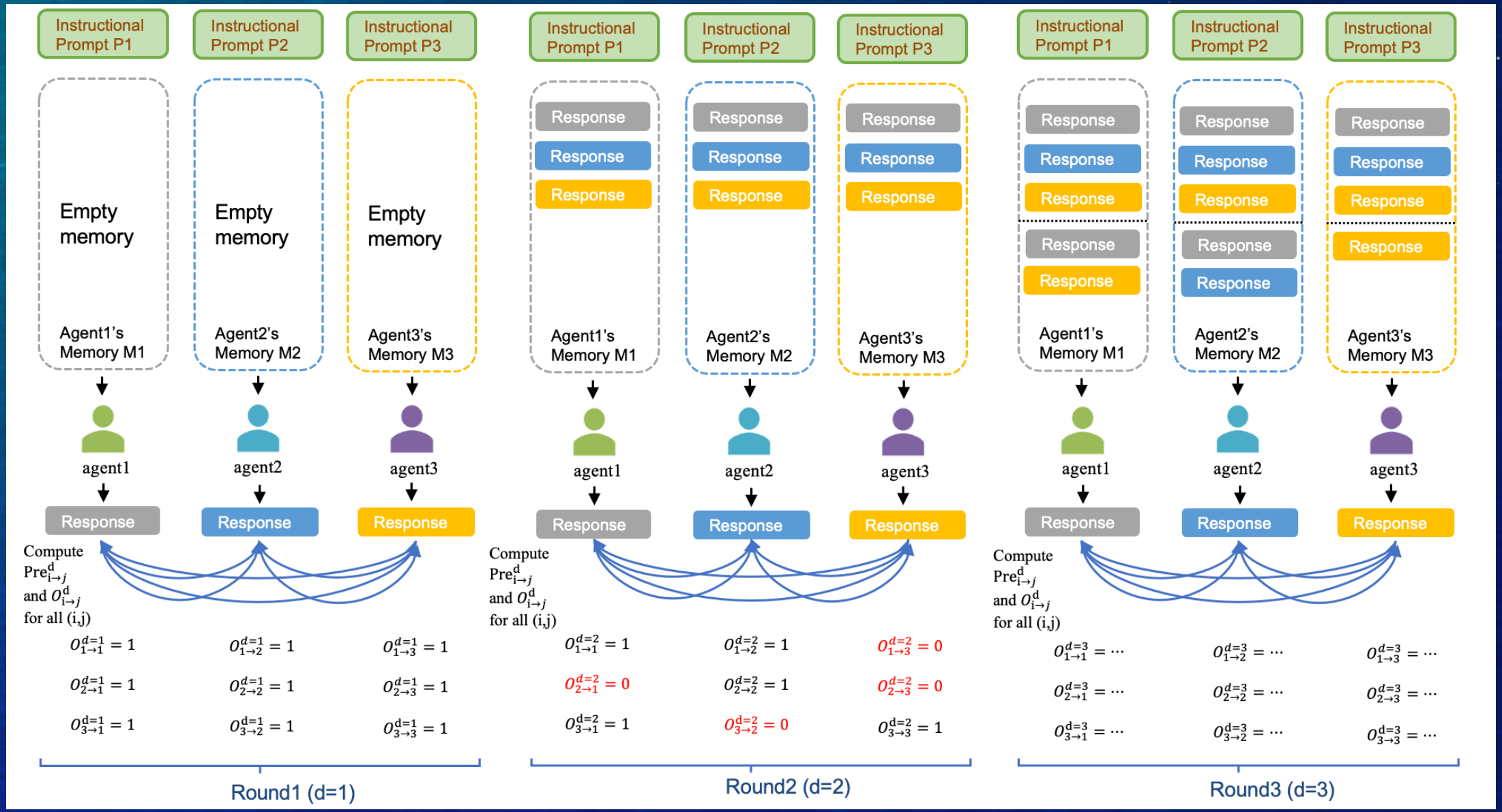
MAD-Logic: Multi-Agent Debate Enhances Symbolic Translation and Reasoning. ICLR, 2026

双阶段辩论：先校正翻译，再融合推理



MAD-Logic: Multi-Agent Debate Enhances Symbolic Translation and Reasoning. ICLR, 2026

选择性记忆更新：按需保留智能体反馈



MAD-Logic: Multi-Agent Debate Enhances Symbolic Translation and Reasoning. ICLR, 2026

稀疏通信与推理多样性兼顾准确率和效率

在效率提升的同时，稀疏通信方法也**优于全局通信**方法。

Model	Method	ProntoQA		ProofWriter		LogicalDeduct	
		Tokens ↓	Acc (%) ↑	Tokens ↓	Acc (%) ↑	Tokens ↓	Acc (%) ↑
GPT-4	SparseMAD	37,784	99.80	41,678	89.50	**43,635**	88.67
	CortexDebate	35,973	99.60	**37,554**	90.83	45,487	92.33
	Ours (w/o sparse)	46,502	99.40	51,358	90.17	54,857	94.00
	Ours (w/ sparse)	**35,854**	**100.00**	42,617	**92.00**	54,171	**94.33**
Claude 3.7	SparseMAD	79,456	99.80	48,190	92.83	**44,245**	99.83
	CortexDebate	68,023	99.80	**41,897**	96.17	45,962	99.67
	Ours (w/o sparse)	106,015	**100.00**	63,105	**97.00**	68,636	99.67
	Ours (w/ sparse)	**52,923**	**100.00**	62,204	96.83	47,121	**100.00**
DeepSeek-V3	SparseMAD	40,200	98.00	**18,527**	92.50	53,257	95.33
	CortexDebate	**35,381**	99.80	19,349	93.00	47,388	99.67
	Ours (w/o sparse)	57,366	99.80	25,059	92.83	70,464	**100.00**
	Ours (w/ sparse)	36,702	**100.00**	24,115	**93.33**	**46,107**	**100.00**

		GPT-4			Claude 3.7 Sonnet			DeepSeek-V3		
SL reasoning	NL reasoning	ProntoQA	ProofWriter	LogiDeduct	ProntoQA	ProofWriter	LogiDeduct	ProntoQA	ProofWriter	LogiDeduct
FOL	COT	97.00%	85.50%	81.67%	99.20%	93.83%	97.67%	98.00%	91.00%	92.00%
SAT+FOL	COT	97.20%	86.17%	93.00%	99.60%	94.00%	99.67%	98.40%	92.50%	99.67%
SAT+FOL+LP	COT	**100.00%**	91.67%	94.00%	**100.00%**	96.17%	**100.00%**	99.60%	92.83%	**100.00%**
SAT+FOL+LP	COT+P&S	**100.00%**	**92.00%**	**94.33%**	**100.00%**	**96.83%**	**100.00%**	**100.00%**	**93.33%**	**100.00%**

参与辩论的符号语言和推理方式种类最多时，能取得最优成绩

MAD-Logic: Multi-Agent Debate Enhances Symbolic Translation and Reasoning. ICLR, 2026

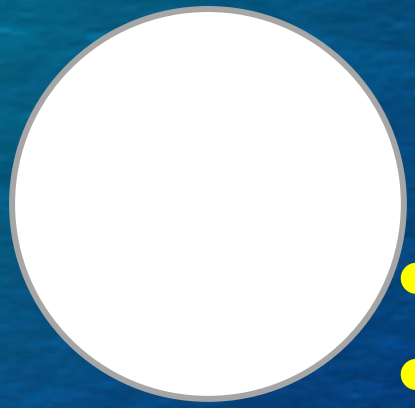
开放世界推理：主动识别并补全缺失前提

- **LLM推理能力短板：**大语言模型在复杂逻辑推理任务中表现不佳，尤其**当推理链较长或需要多步演绎时，**常出现**推理断裂**导致错误；
- **封闭世界假设的局限：**现有方法默认**所有必要前提均已显式提供，**无法处理真实场景中**信息不完整、常识缺失的开放世界问题；**
- **开放世界的迫切需求：**现实决策常依赖**隐含知识，**如法律条文、科学定律等，亟需能**主动识别并补全缺失前提的智能符号推理框架。**

Open-World Logical Reasoning. ICML, 2026

开放世界推理要求从被动演绎转向主动探索

传统封闭世界逻辑推理



所有给定前提

需判定的结论

- 被动推理 v.s. 主动推理
- 封闭世界假设所有前提已知，推理仅为符号推演 v.s.开放世界承认信息不完备，需主动探索缺失前提
- $\Gamma \not\models \varphi$ v.s. 找一个最小命题集 Δ 使得 $\Gamma \cup \Delta \models \varphi$

提出的开放世界逻辑推理

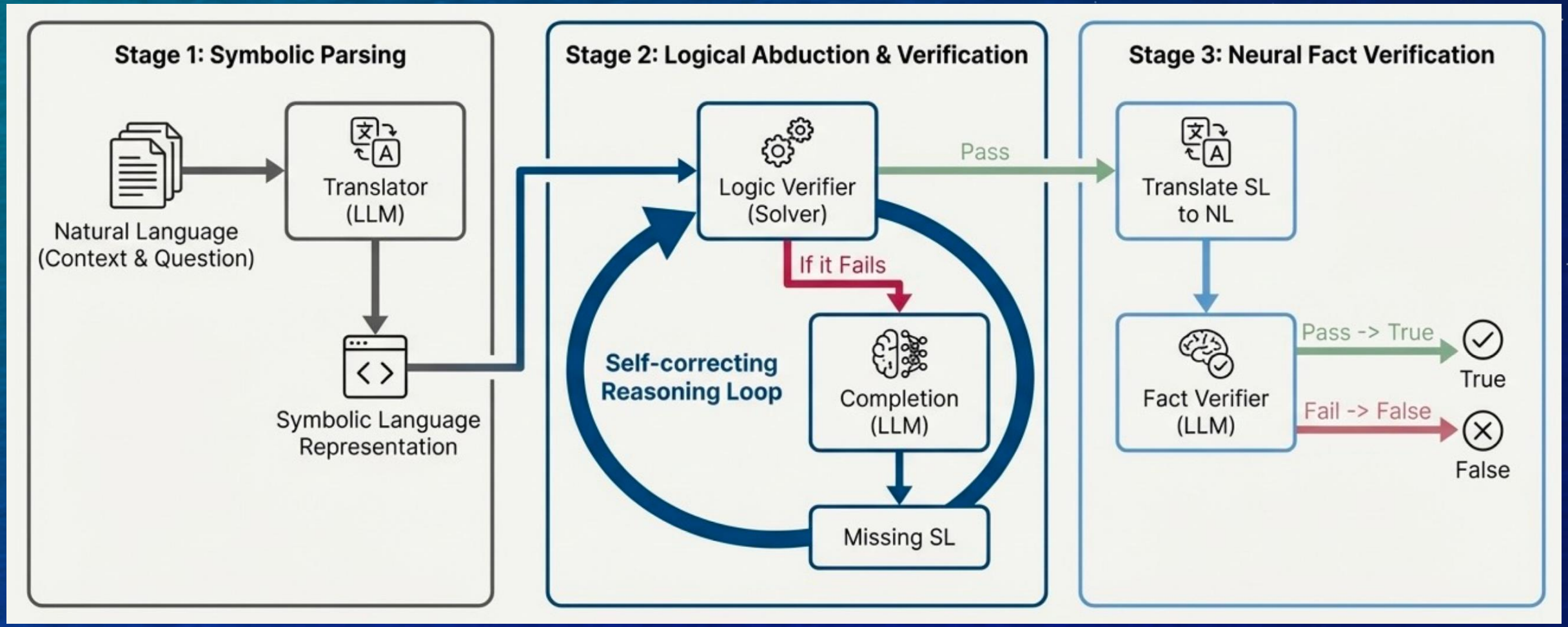


主动寻找缺失前提

需判定的结论

Open-World Logical Reasoning. ICML, 2026

双重验证：前提补全须满足逻辑与事实约束



神经符号翻译：
从自然语言到符号语言

基于逻辑求解器的符号语言补全
（神经符号验证）

转化为自然语言并验证
（自然语言验证）

Open-World Logical Reasoning. ICML, 2026

常识前提补全恢复因知识缺失而中断的推理

Input: Context: “The nail is made of iron.” Question: “Can a nail conduct electricity?”

Initial Verification — FAILED

Logic: MadeOf(nail, iron) ⊭ ConductsElectricity(nail)

↓

Completion & Retry 1 — FAILED

Missing SL Generated: ConductsElectricity(metal)

Logic Check: MadeOf(nail, iron), ConductsElectricity(metal) ⊭ ConductsElectricity(nail)

↓

Completion & Retry 2 — PASSED

Missing SL Generated: IsMetal(iron)

Logic Check: MadeOf(nail, iron), ConductsElectricity(metal), IsMetal(iron) ⊨ ConductsElectricity(nail)

→

Fact Verification

- Checks: “Metal can conduct electricity.” → **TRUE**
- Checks: “Iron is Metal.” → **TRUE**

→ **True**

最小前提补全：减少前提数量以收紧误差界

Theorem 3.1 (PAC-Bayes Bound). *For any prior* $P(\Delta)$ *and posterior* $Q(\Delta)$ *over the missing premises* Δ *and any* $\delta \in (0,1)$, *with probability at least* $1-\delta$,

$$R_{\mathcal{D}}(Q(\Delta)) \leq \hat{R}_S(Q(\Delta)) + \sqrt{\frac{\mathrm{KL}(Q(\Delta)\|P(\Delta)) + \log\frac{1}{\delta}}{2(n-1)}}.$$

Theorem 3.2 (Monotonicity). *Fix* $\alpha > 0$, n, *and* δ. *Consider* Q_s *and* Q_l *such that* $\mathbb{E}_{Q_s}[|\Delta|] \leq \mathbb{E}_{Q_l}[|\Delta|]$,

$$\mathbb{E}_{Q_l}[|\Delta|] - \mathbb{E}_{Q_s}[|\Delta|] \geq \frac{\mathbb{E}_{Q_s}[\log Q_s(\Delta)] - \mathbb{E}_{Q_l}[\log Q_l(\Delta)]}{\alpha},$$

then the PAC-Bayes upper bound for Q_s *is smaller than* Q_l.

PAC-贝叶斯泛化误差界

泛化误差界

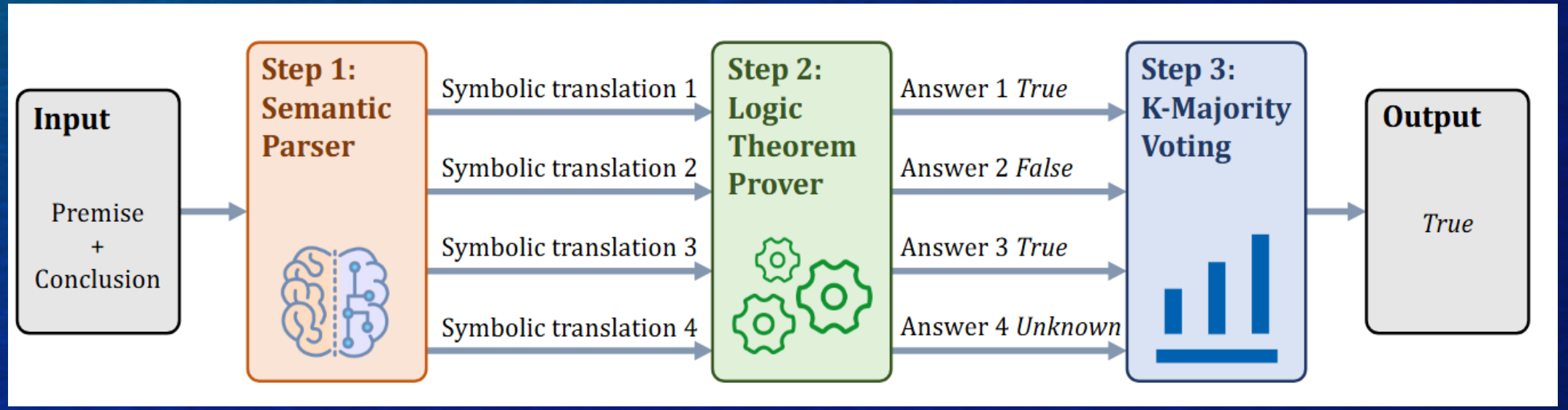
随逻辑前提数量增加

而增加

Open-World Logical Reasoning. ICML, 2026

神经符号分工：大模型翻译，求解器推理

1. 提示大语言模型将自然语言问题**转译**为符号公式。
2. 调用相应的**逻辑求解器**来推断答案。
3. 使用集成方法生成最终答案，例如多数投票。



Open-World Logical Reasoning. ICML, 2026

Logic-LM：支持多符号语言的形式化和推理

- **问题形式化器：**

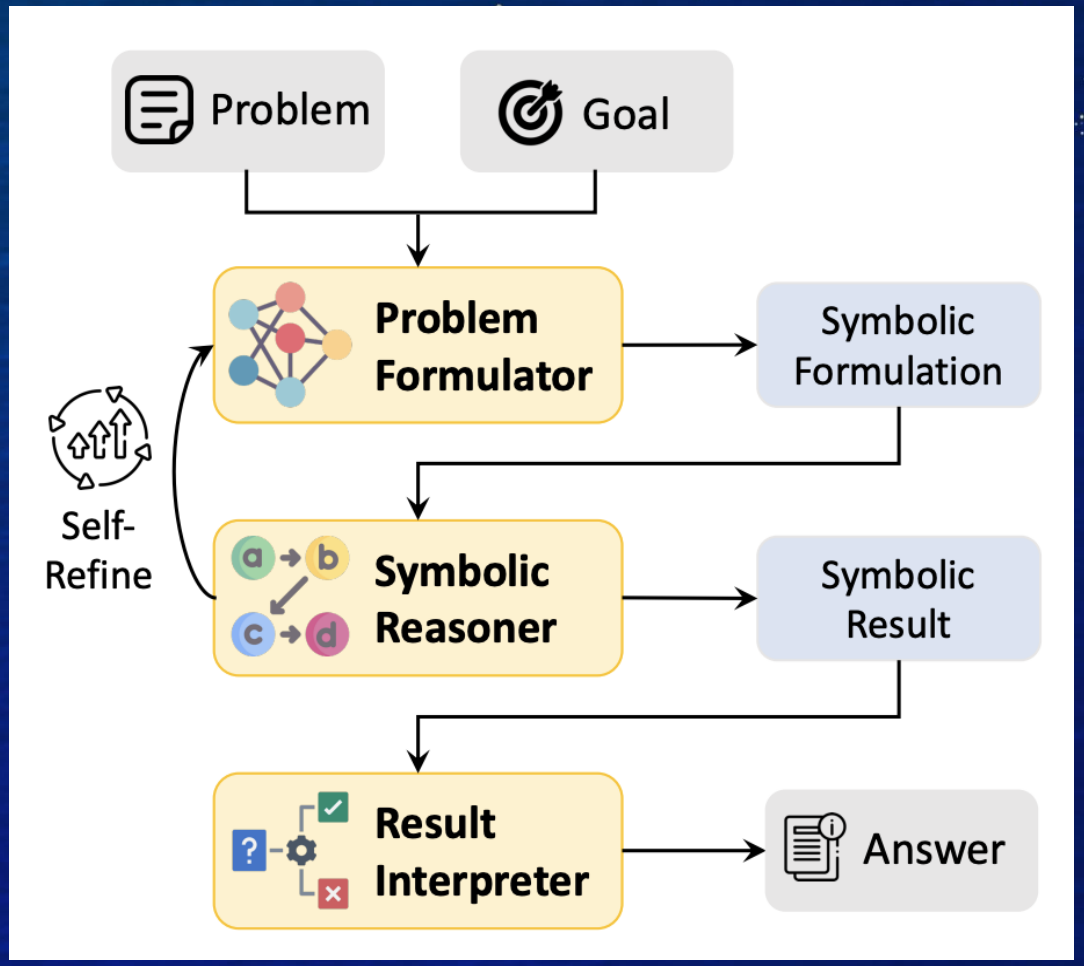
  通过情境学习，用大语言模型为输入问题生成符号表示。

- **符号推理器：**

  对形式化后的问题执行逻辑推断。

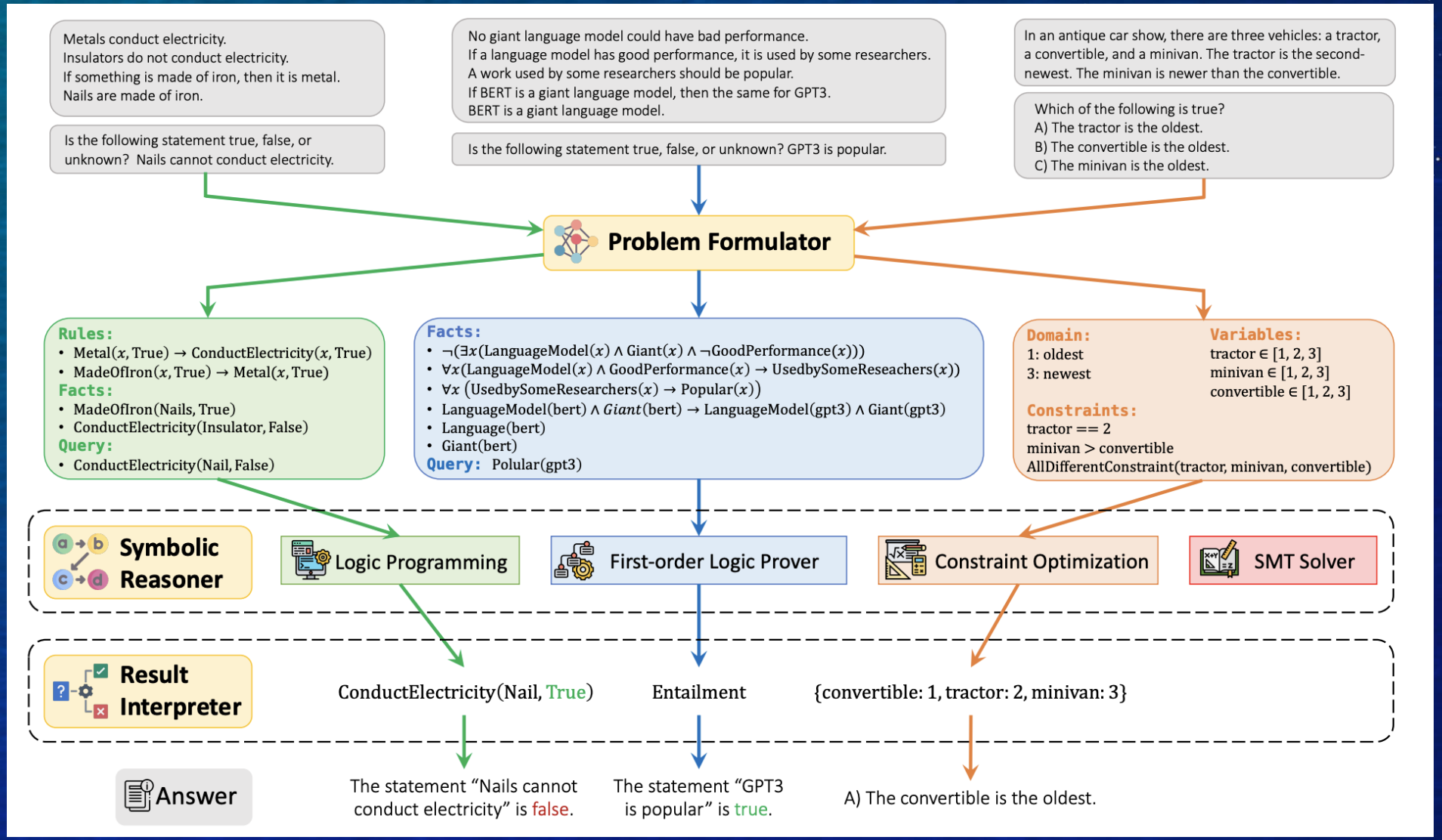
- **结果解释器：**

  对符号答案进行解释。



Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In Findings of EMNLP, 2023.

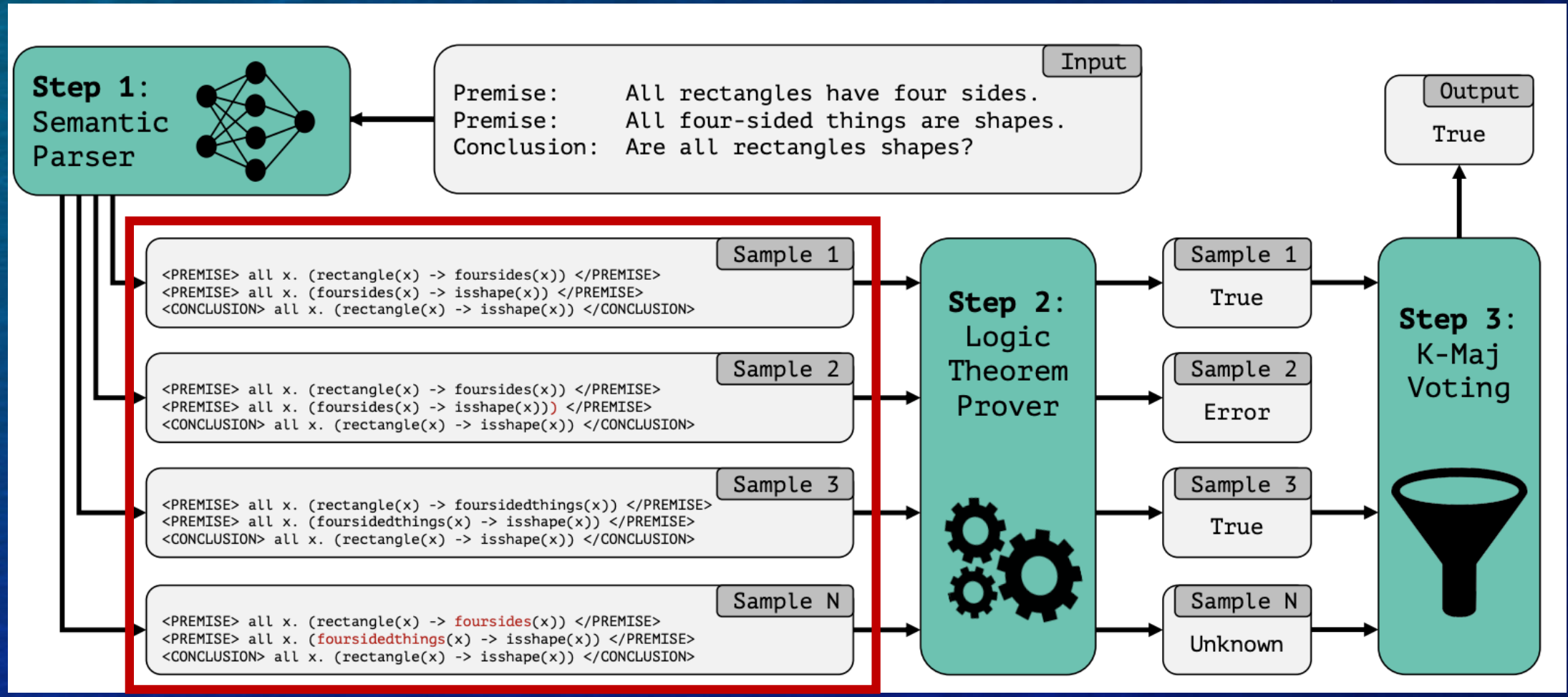
Logic-LM：支持多符号语言的形式化和推理



Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In Findings of EMNLP, 2023.

LINC：基于多数投票的一阶逻辑翻译和求解

为了减少翻译错误



LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In EMNLP, 2023

Logic.py：逻辑约束转化是复杂谜题推理的瓶颈

- **核心问题：**大语言模型（LLM）难以处理逻辑任务（如斑马谜题、一阶逻辑问题）
  - 基准准确率：Llama 3.1 70B → 24.9% (ZebraLogicBench)
  - 根本原因：自然语言（NL）向逻辑约束的转化效果不佳
- **挑战：**LLM 缺乏严谨的推理能力；纯符号方法无法应对自然语言的歧义性
- **目标：**连接 LLM（自然语言理解）与约束求解器（严谨推理）

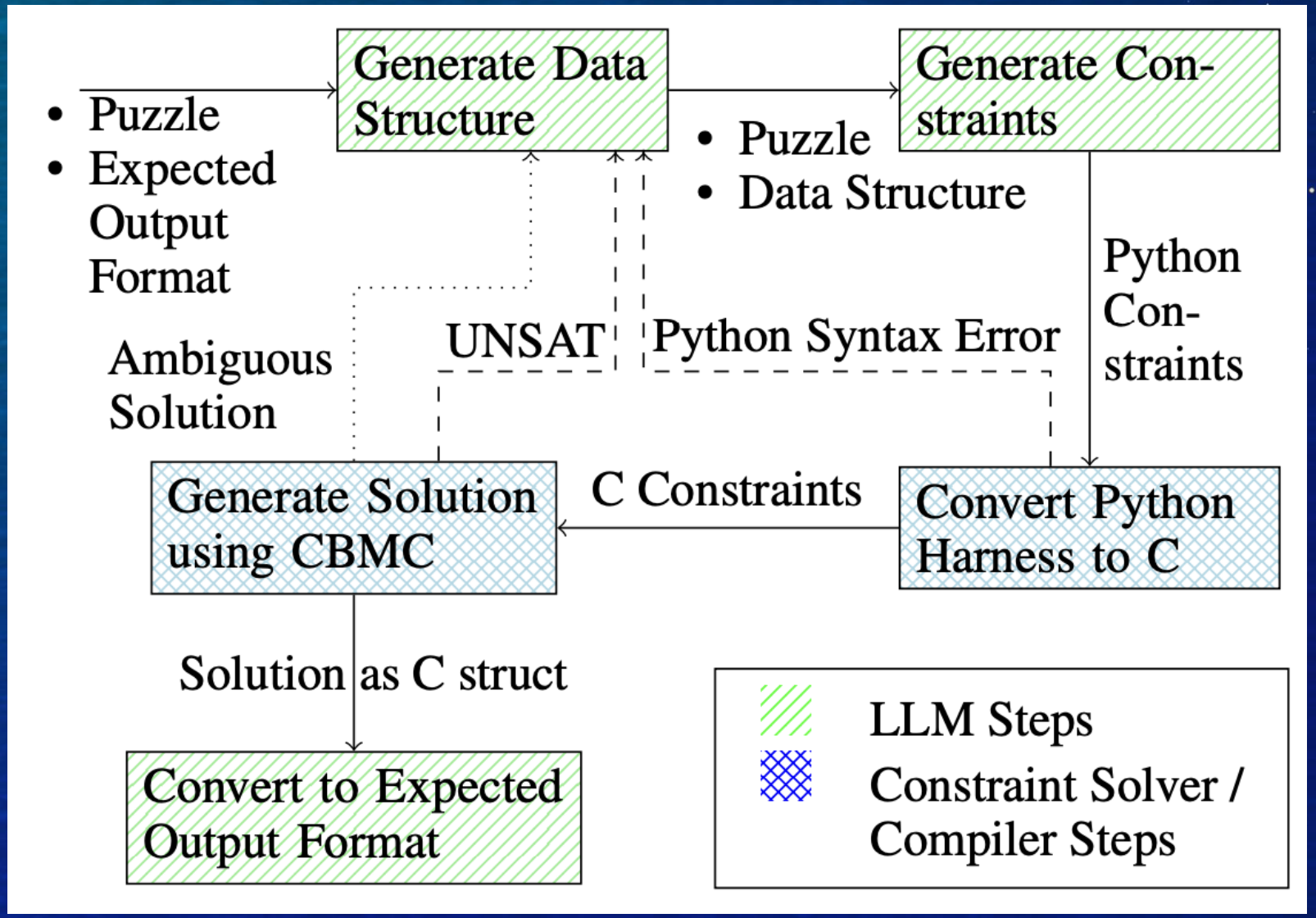
Logic.py: Bridging the Gap between LLMs and Constraint Solvers. In NeurIPS, 2025.

Logic.py：用可读逻辑语言衔接理解与严谨求解

- **核心组件：**
  - Logic.py：一种用于将自然语言转换为逻辑表示、且便于人类阅读的领域特定语言（DSL）
  - Polymath Agent：连接 “大语言模型 → Logic.py → 约束求解器（CBMC/Z3）” 的智能体
- **设计标准：**
  - 人类可读性、鲁棒性、简洁性和表达能力

Logic.py: Bridging the Gap between LLMs and Constraint Solvers. In NeurIPS, 2025.

Polymath：利用求解反馈修复错误约束



Logic.py: Bridging the Gap between LLMs and Constraint Solvers. In NeurIPS, 2025.

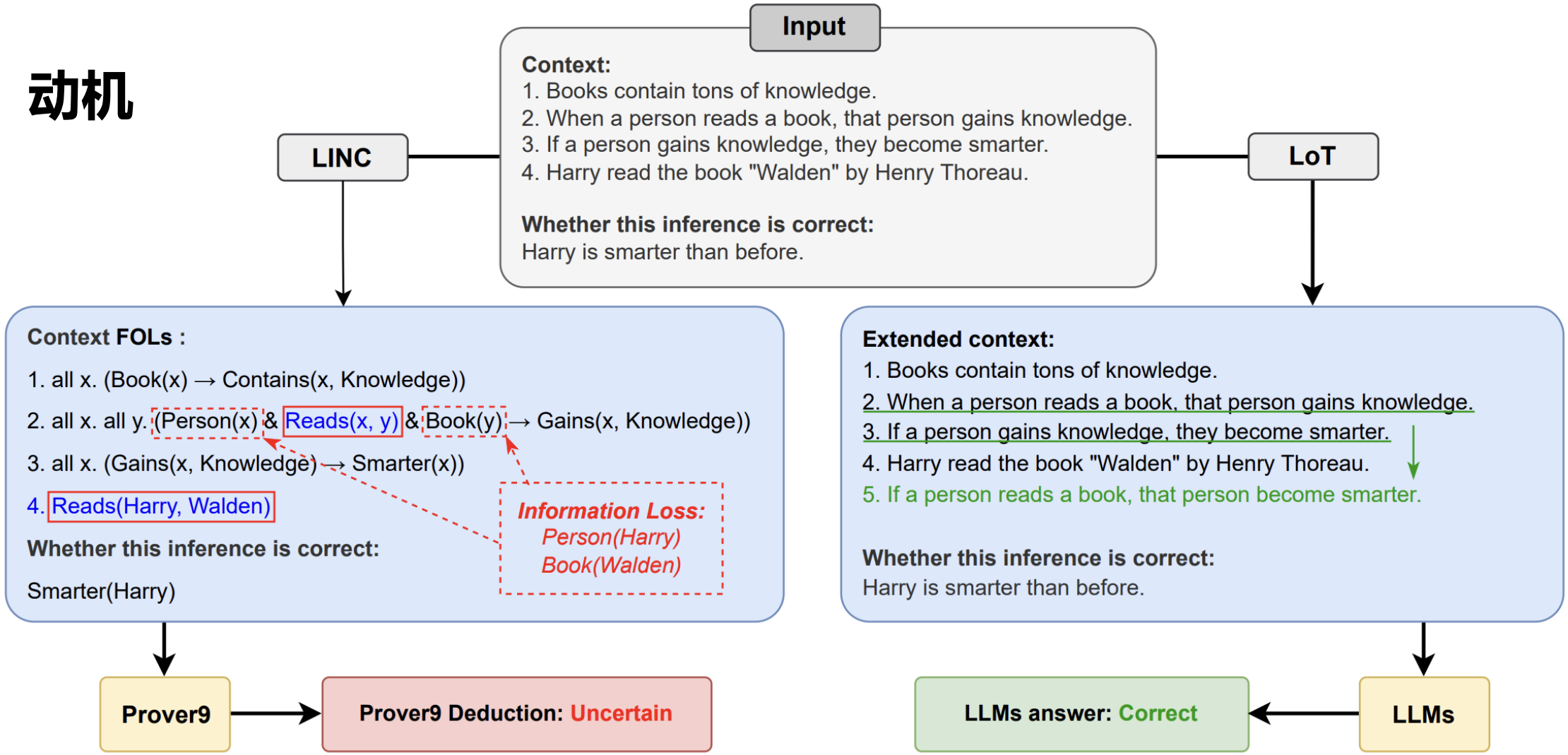
Logic.py：约束求解大幅提升复杂谜题准确率

Model	Puzzle Accuracy					Cell Accuracy
	All	Small	Medium	Large	XL	
grok-3-mini-fast-beta-high	92.6	98.75	96.43	93.5	76.5	94.63
o3-mini-2025-01-31-high	91.7	99.69	97.14	87.5	75.5	95.7
Polymath Meta-Llama-3.1-70B-Instruct	90.7	93.75	93.21	90	83	92.56
o3-mini-2025-01-31-medium	88.9	99.69	97.86	88	60	90.41
o1-2024-12-17	81	97.19	92.14	78	42.5	78.74
grok-3-mini-fast-beta-low	80.7	98.75	96.43	77	33.5	84.22
deepseek-R1	78.7	98.44	95.71	73.5	28.5	80.54
Polymath gpt-4o	77.8	91.88	79.64	75.5	55	76.34
Polymath claude-3-5-sonnet-20241022	76.9	85.31	78.57	76	62	80.58
o3-mini-2025-01-31-low	74.8	99.38	91.07	64.5	23	72.6
o1-preview-2024-09-12	71.4	98.12	88.21	59.5	17	75.14
claude-3-5-sonnet-20241022	36.2	84.69	28.93	4	1	54.27
Llama-3.1-405B-Inst-fp8@together	32.6	81.25	22.5	1.5	0	45.8
gpt-4o-2024-08-06	31.7	80	19.64	2.5	0.5	50.34
gemini-1.5-pro-exp-0827	30.5	75.31	20.71	3	0	50.84
Mistral-Large-2	29	75.94	15	2.5	0	47.64
Qwen2.5-72B-Instruct	26.6	72.5	12.14	0	0	40.92
Meta-Llama-3.1-70B-Instruct	24.9	67.81	10.36	1.5	0	27.98

Logic.py: Bridging the Gap between LLMs and Constraint Solvers. In NeurIPS, 2025.

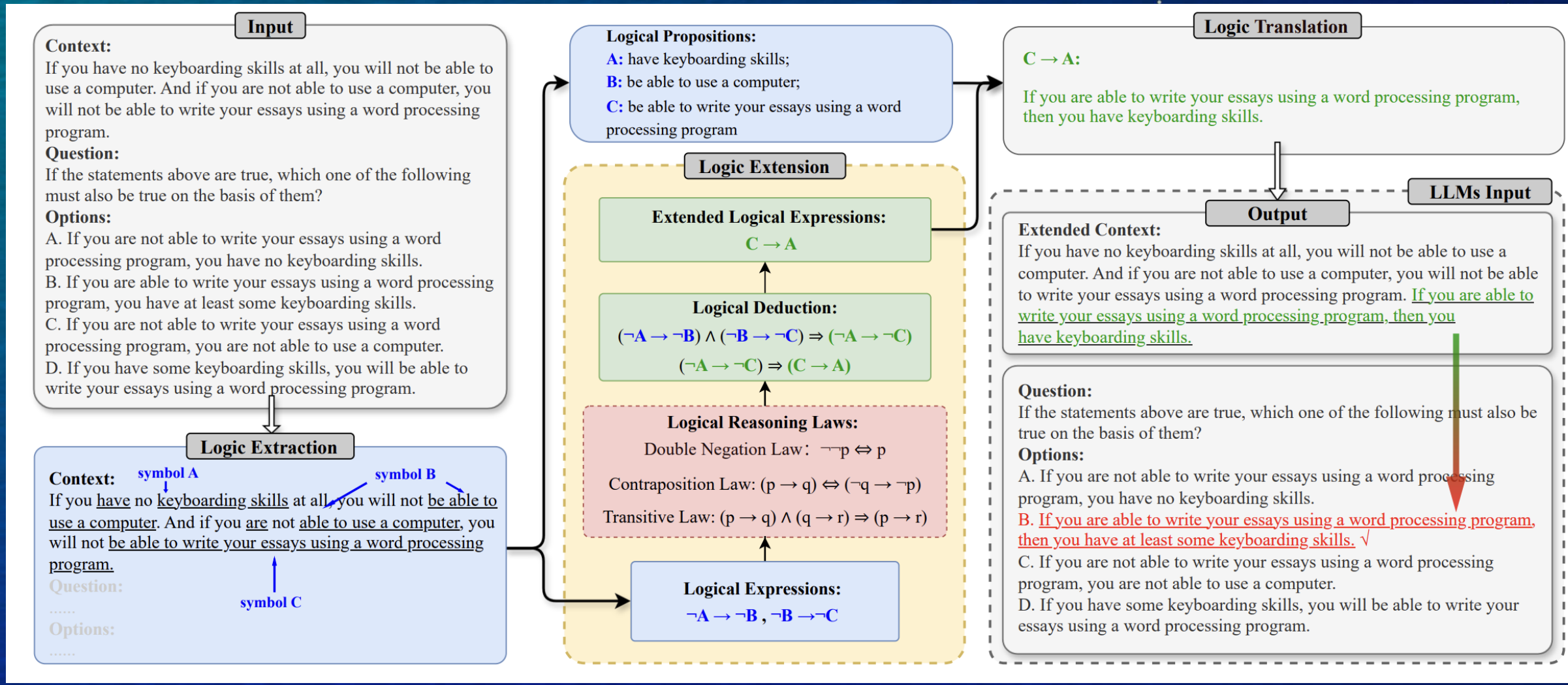
LoT：保留原始语境以缓解符号化信息损失

动机



Logic-of-thought: Injecting logic into contexts for full reasoning in large language models. In NAACL, 2025.

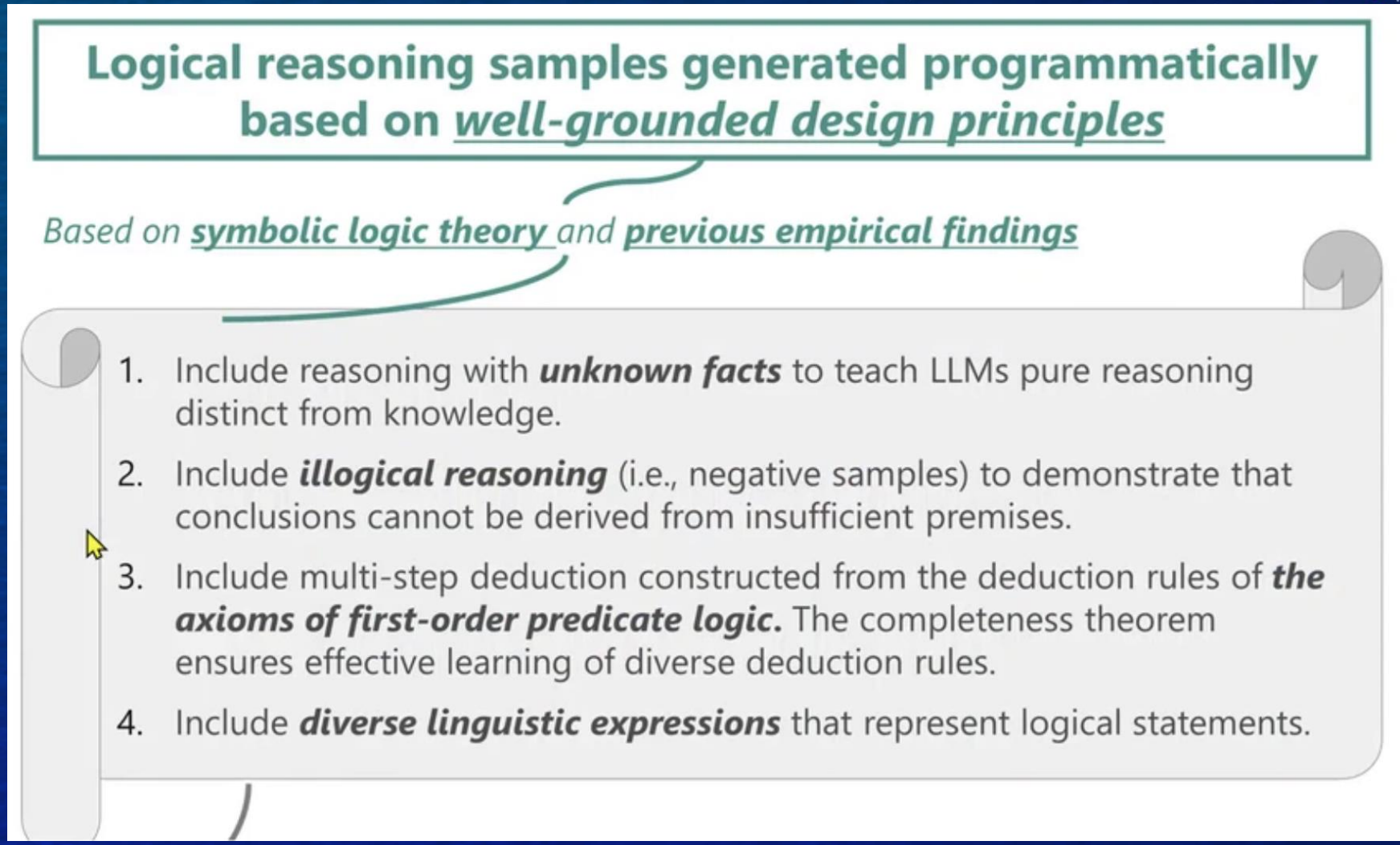
LoT：显式补充隐含逻辑关系以辅助推理



Logic-of-thought: Injecting logic into contexts for full reasoning in large language models. In NAACL, 2025.

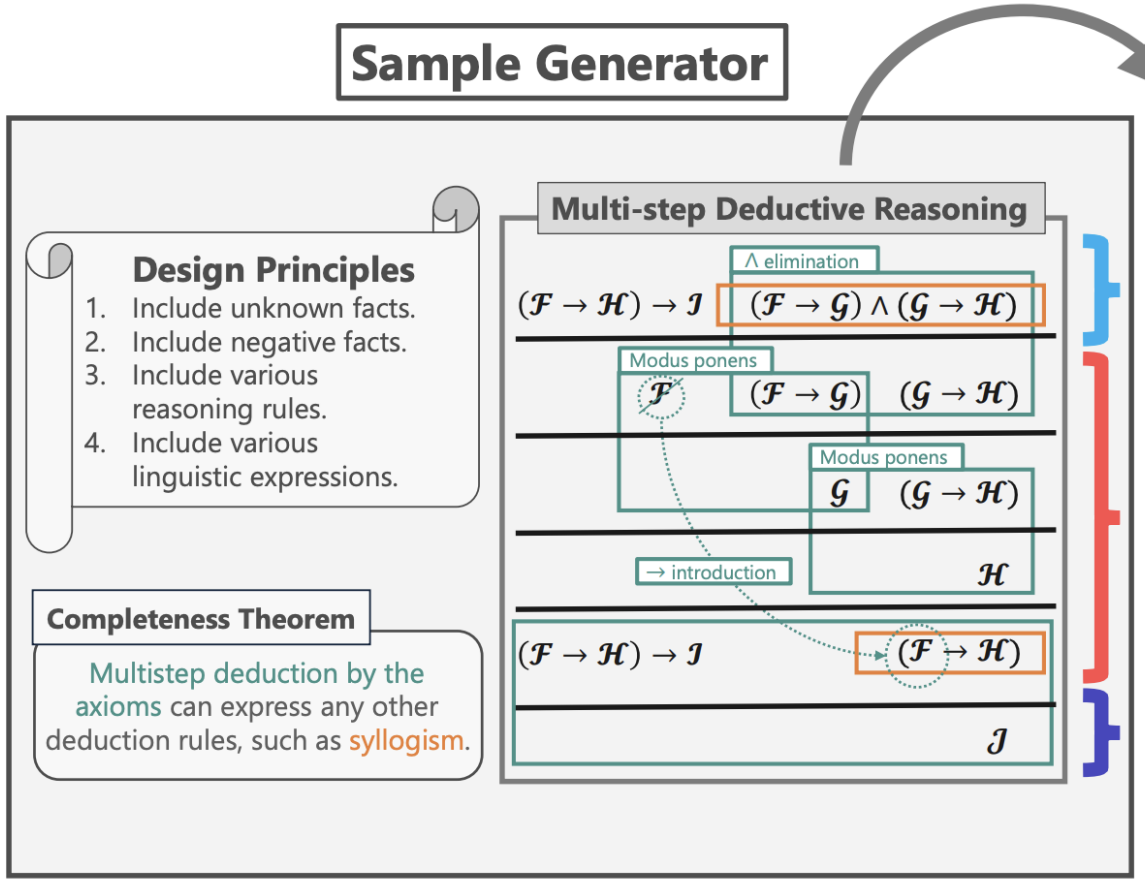
ALT：多样化逻辑样本支撑推理能力泛化

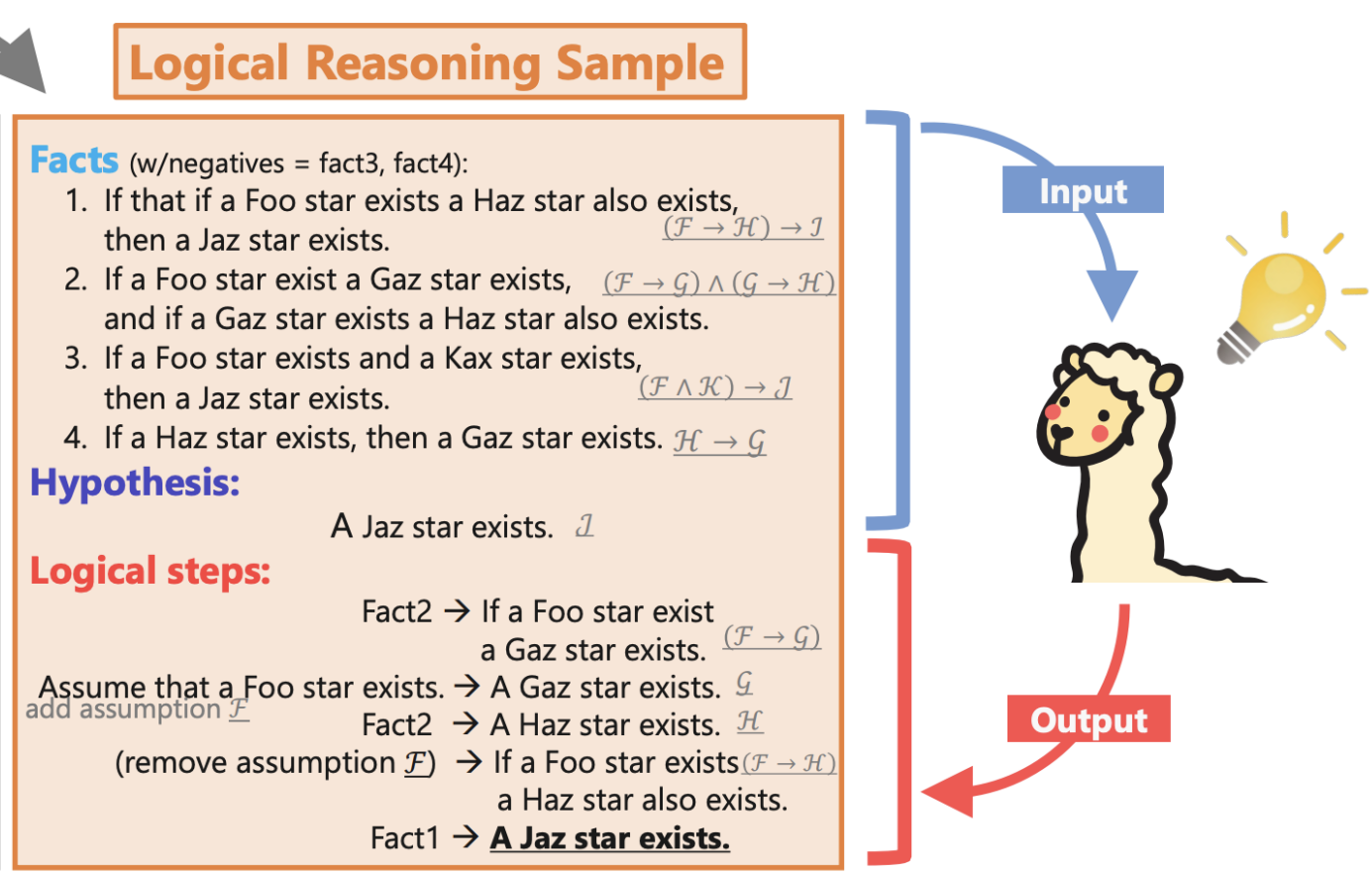
- 动机：大语言模型（LLM）推理能力较弱，可能源于预训练语料库中缺乏高质量的推理样本。
- 为训练准备高质量的推理样本。



Enhancing reasoning capabilities of LLMs via principled synthetic logic corpus. In NeurIPS, 2024

ALT：公理组合使合成训练覆盖多样推理规则

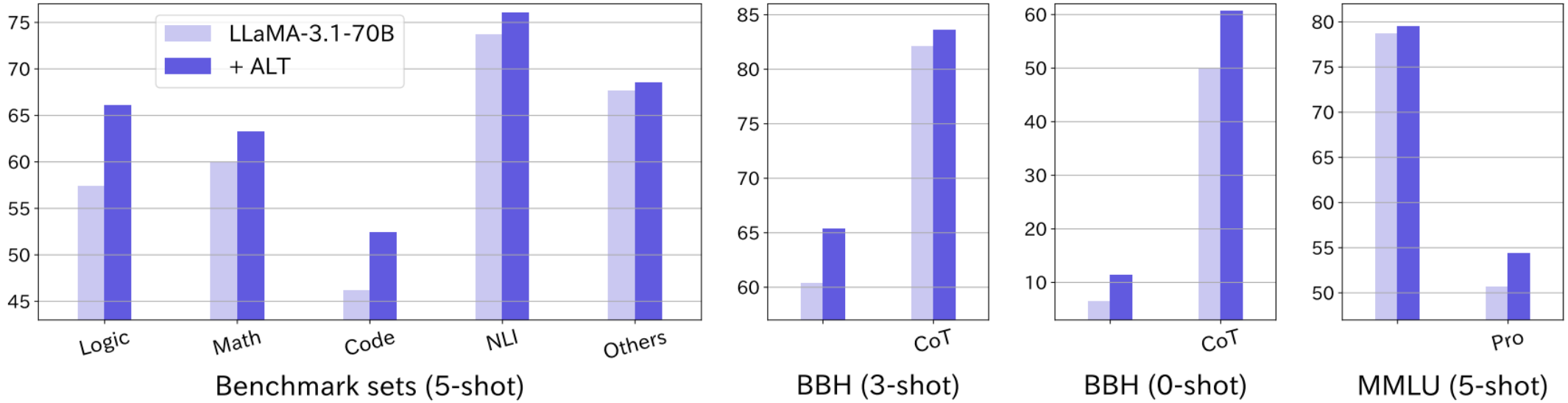




Enhancing reasoning capabilities of LLMs via principled synthetic logic corpus. In NeurIPS, 2024

ALT：演绎训练可泛化到溯因与通用推理

FLD×2 substantially enhances the reasoning capabilities of SOTA LLMs



Logic

- Up to +30 points
- Enhance abduction → generalizability

Abduction

	bAbiD	FOLIO	LogicNLI	RobustLR	AR-LSAT	LogiQA	ReClor	AbductionR	ART
LLaMA-3.1-70B	**83.8**$_{\pm 1.2}$	58.9$_{\pm 1.6}$	34.9$_{\pm 1.1}$	49.6$_{\pm 0.9}$	21.5$_{\pm 1.0}$	64.3$_{\pm 1.2}$	33.7$_{\pm 0.7}$	84.0$_{\pm 0.7}$	85.4$_{\pm 0.9}$
⊕FLD×2	83.5$_{\pm 0.5}$	**66.7**$_{\pm 0.6}$	**50.9**$_{\pm 0.5}$	**81.6**$_{\pm 0.3}$	**25.0**$_{\pm 0.4}$	**69.4**$_{\pm 0.5}$	**36.3**$_{\pm 0.3}$	**95.7**$_{\pm 0.2}$	**85.5**$_{\pm 0.4}$

Enhancing reasoning capabilities of LLMs via principled synthetic logic corpus. In NeurIPS, 2024

ALT：逻辑训练向数学与代码任务迁移

Math

- Up to +7 points
- Predicate logic is a prerequisite for math

	GSM8k			MATH	MathQA
		CoT	CoT (0-shot)	-	-
LLaMA-3.1-70B	$80.9_{\pm 1.1}$	$75.2_{\pm 1.2}$	$65.4_{\pm 1.3}$	$23.7_{\pm 0.6}$	$55.0_{\pm 0.9}$
⊕**FLD**$^{\times 2}$	**83.3**$_{\pm 0.4}$	**80.4**$_{\pm 0.4}$	**73.0**$_{\pm 0.5}$	**24.4**$_{\pm 0.2}$	**55.4**$_{\pm 0.4}$

Code

- Up to +10 points
- Consistent with other studies

	HumanEval	MBPP	MBPP+	MultiPL-E (cpp)	MultiPL-E (go)
LLaMA-3.1-70B	32.3	43.4	48.7	29.8	76.6
⊕**FLD**$^{\times 2}$	**42.6**	**49.5**	**52.5**	**38.7**	**78.6**

NLI

- Up to +6 points
- Integrate knowledge and reasoning

	HELP	MNLI	RTE	SNLI
LLaMA-3.1-70B	$45.8_{\pm 0.5}$	$82.2_{\pm 0.4}$	$84.0_{\pm 0.7}$	**82.6**$_{\pm 0.4}$
⊕**FLD**$^{\times 2}$	**51.3**$_{\pm 0.2}$	**83.7**$_{\pm 0.2}$	**87.2**$_{\pm 0.3}$	$82.3_{\pm 0.2}$

Others

- Up to + 2 points
- FLD$^{\times 2}$ does not teach knowledge or complex procedures

	CommonsenseQA	HellaSwag	SQuAD	WinoGrande	ARCe	ARCc	GPQA	OpenBookQA	SciQ
LLaMA-3.1-70B	$81.2_{\pm 1.1}$	$69.2_{\pm 0.5}$	$38.5_{\pm 0.0}$	$85.6_{\pm 1.0}$	$89.1_{\pm 0.6}$	$65.3_{\pm 1.4}$	**40.7**$_{\pm 1.4}$	$41.4_{\pm 0.7}$	**98.5**$_{\pm 0.4}$
⊕**FLD**$^{\times 2}$	**82.5**$_{\pm 0.4}$	**69.6**$_{\pm 0.2}$	**40.1**$_{\pm 0.0}$	**86.1**$_{\pm 0.4}$	**89.4**$_{\pm 0.3}$	**66.7**$_{\pm 0.6}$	$40.6_{\pm 0.6}$	**42.8**$_{\pm 0.3}$	**98.5**$_{\pm 0.2}$

© Hitachi, L

Enhancing reasoning capabilities of LLMs via principled synthetic logic corpus. In NeurIPS, 2024

SynLogic：可验证逻辑数据支撑强化学习

核心差距

- 开源 LLM RL 训练主要侧重于数学/编码，缺乏通用逻辑推理的数据
- 现有逻辑数据集：任务很少（1-5），不适合训练，难度固定

关键动机

- 通过构建**大规模、可验证、难度可调**的数据集来提高 LLM 的通用推理能力，解决高质量逻辑训练数据的短缺问题。

SynLogic: Synthesizing Verifiable Reasoning Data at Scale for Learning Logical Reasoning and Beyond. In NeurIPS, 2025.

SynLogic：兼顾多任务训练与难度控制

Dataset	Tasks	Trainable	Adjustable Difficulty
BBH [Suzgun et al., 2022]	23	✗	✗
Zebra Logic [Lin et al., 2024]	1	✗	✓
KOR-Bench [Ma et al., 2024]	5*	✗	✗
K&K [Xie et al., 2025a]	1	✓	✓
BBEH [Kazemi et al., 2025]	23	✗	✗
SYNLOGIC	35	✓	✓

SynLogic: Synthesizing Verifiable Reasoning Data at Scale for Learning Logical Reasoning and Beyond. In NeurIPS, 2025.

LogicTree：简单模板难以支撑复杂多步推理

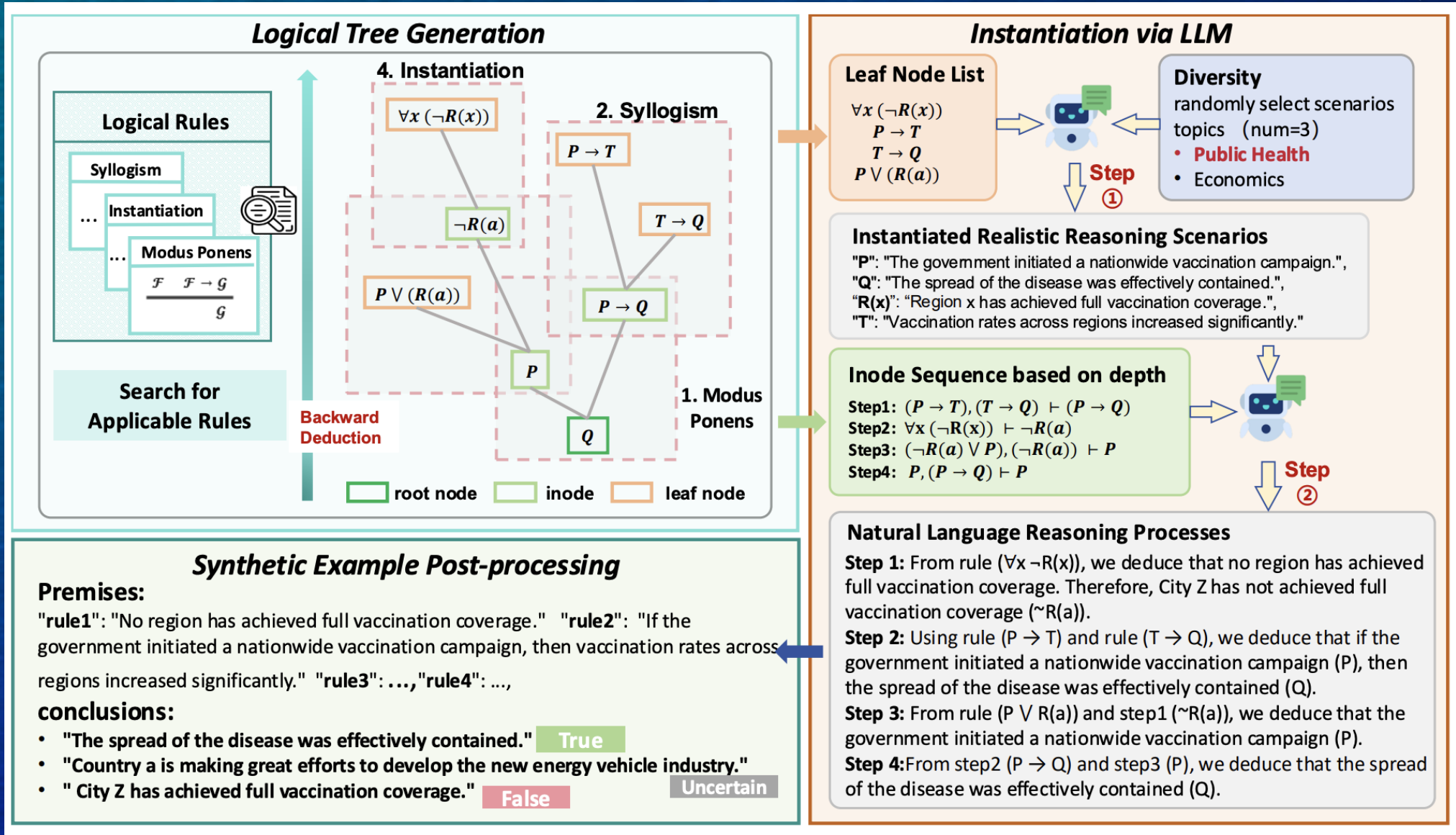
由于缺乏高质量的训练数据，LLM很难进行**多步骤的复杂逻辑推理**

现有方法的局限性

- **复杂性不够：**推理模式简单、规则类型少、步骤浅
- **现实世界的实例化较差：**实体断开，泛化能力弱

LogicTree: Improving Complex Reasoning of LLMs via Instantiated Multi-step Synthetic Logical Data. In NeurIPS, 2025.

LogicTree：场景实例化促进复杂推理泛化

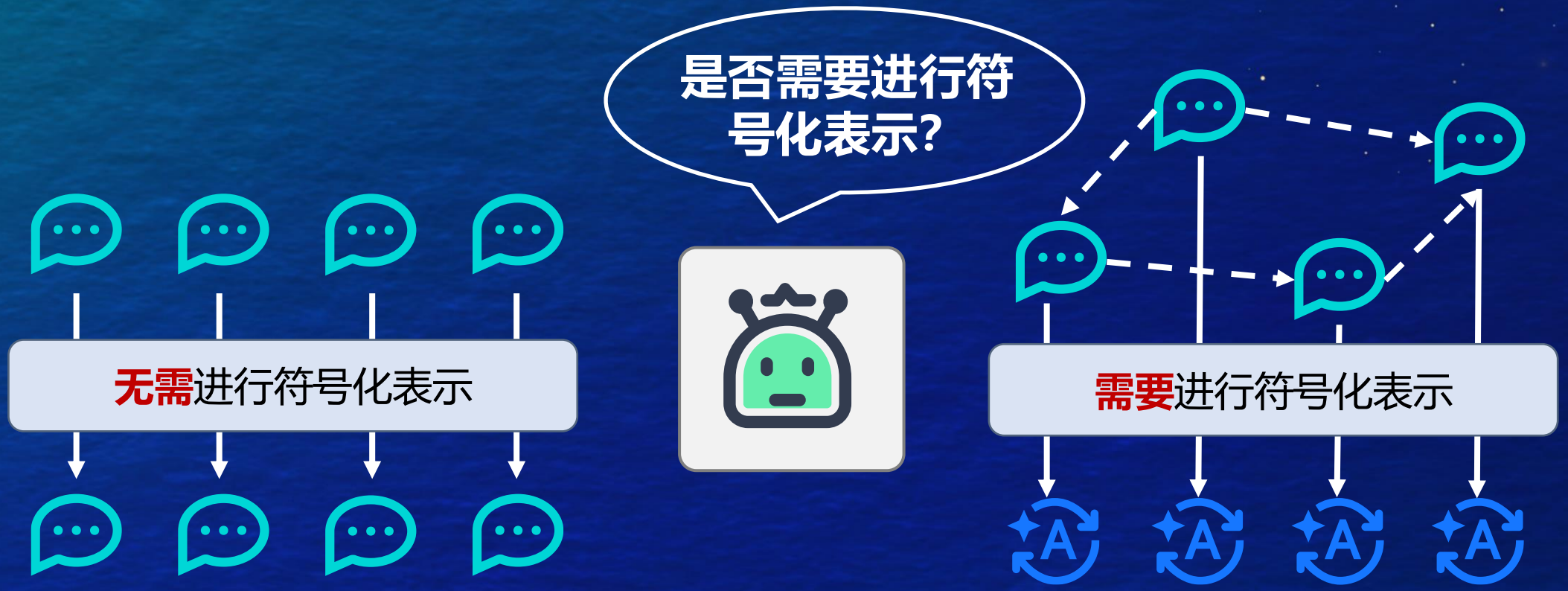


LogicTree: Improving Complex Reasoning of LLMs via Instantiated Multi-step Synthetic Logical Data. In NeurIPS, 2025.

未来研究方向1：自适应符号语言表示

自然语言　符号语言

- 哪些自然语言信息需要进行符号化表示？



未来研究方向2：翻译准确率和信息损失

符号语言表示

- 如何对其进行表示（本质上就是一种翻译）？
- 如何提高这种翻译的质量？如何应对翻译带来的信息损失?



未来研究方向3：扩展至模态逻辑，以表达不确定性

- 我们重点研究**模态逻辑**和**认知逻辑**，以评估并提升大语言模型处理更复杂、更多样化推理任务的能力
- 模态逻辑通过引入 **“必然”（□）**和 **“可能”（◇）**两个算子来扩展命题逻辑，分别表示必然性和可能性
- 大语言模型的**高阶逻辑推理**
- 对属性和函数进行推理

谢谢观看