



華科技大學

人工智能与自动化学院

School of Artificial Intelligence and Automation

张同学

2026 年 7 月

博士学位论文答辩 / PhD Dissertation Defense

面向可信人工智能的 结构化推理方法

Structured Reasoning for Trustworthy
Artificial Intelligence

汇报结构

01

问题与目标

可信系统不仅要给出答案，还要提供可检查的中间证据。

02

方法

通过任务分解、局部求解和一致性验证组织推理过程。

03

结果与讨论

在有限的额外计算成本下，显著提升错误恢复能力。

问题与目标

无法检查的过程，不足以支撑可信结论。

为什么需要结构化推理？

从答案到证据

- 大模型能够生成流畅答案，但中间过程往往不可检查。
- 学术与工程场景更关注可追溯性、稳定性与错误边界。
- 复杂任务需要被拆解为可以独立验证的推理单元。

设计目标

可验证

输入、输出与检查均明确。

可恢复

失败后只重算受影响的局部步骤。

三个设计原则

01 / DECOMPOSE

可分解

将长链问题拆分成短而明确的子任务。

02 / VERIFY

可验证

每一步都产生可以独立检查的中间状态。

03 / RECOVER

可恢复

检测失败后，只重算受影响的局部步骤。

共同约束

模块之间只交换显式状态；验证器负责给出通过、拒绝或局部修正信号。

Item 建立阅读顺序

ITEMIZE / 并列信息

- **结论先行**
每项先给出观点，再补充一行依据。
- **句式平行**
同一层级保持相近的信息颗粒度。
- **层级克制**
 - 次级条目补充条件或例外。
 - 三级仅用于必要注释。

ENUMERATE / 顺序步骤

- 01 定义输入**
问题、状态与约束保持明确。
- 02 执行任务**
 - a 生成候选
 - b 保存中间状态
- 03 检查输出**
失败时只进入局部恢复。

Box 用标题带建立语义边界

蓝色 / 核心信息与结论

- 窄蓝色标题带负责命名主语义区域。
- 淡紫灰正文面承载解释、比较与 Future Work。

粉色 / 定义、公式与局部强调

- 粉色标题带用于定义、公式与次级语义区域。
- 真正的风险与失败仍然只用红色文字表达。

方法

先分解，再验证；失败只回滚局部。

结构化推理框架

模块化执行与局部恢复



深灰表示前向主流程；红色只表示验证失败后的局部反馈。

目标函数与局部恢复

联合优化目标

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{verify}} + \beta \mathbb{E}[\mathbf{C}].$$

$\mathcal{L}_{\text{task}}$ 最终任务误差

$\mathcal{L}_{\text{verify}}$ 中间状态验证误差

$\mathbb{E}[\mathbf{C}]$ 平均计算成本

局部恢复性

若依赖图无环且节点验证器完备，单点失败只需重算后继闭包，不影响独立分支。

失败 $\rightarrow$ 后继闭包 $\rightarrow$ 重算

最小实现与接口契约

```
1 def reason(problem, modules):
2     plan, solve, verify = modules
3     state = {}
4     for step in plan(problem):
5         result = solve(step, state)
6         if not verify(step, result, state):
7             result = solve(step, state, revise=True)
8         state[step.id] = result
9     return result
```

接口契约

输入 问题与已验证状态
验证 通过或拒绝

输出 可复核的局部结果
失败 仅修正当前依赖分支

结果与讨论

关键结果：在有限额外步骤下，可恢复率达到 73.5%。

实验结果

四种推理策略对比

方法	正确率 ↑	平均步数 ↓	可恢复率 ↑
Direct	71.4	1.0	0.0
Chain-of-Thought	78.9	6.8	12.3
Self-Consistency	82.1	34.0	18.7
Ours	86.7	9.4	73.5

示例数据仅用于验证信息密度、表格层级与强调语义。

关键结果

86.7%

正确率领先对照。

73.5%

失败可定位并恢复。

结论

01

结构化表示

让推理过程可以被检查、追踪并局部修复。

02

显式验证

在合理计算成本下，提高系统的可靠性与恢复能力。

03

下一步

研究弱监督条件下的自动验证器与动态任务分解。

华中科技大学

感谢聆听

Questions & Discussion

张同学 · 人工智能与自动化学院