



中国科学院大学

University of Chinese Academy of Sciences

硕士学位论文

基于图谱补全和过程验证的因果推理研究与应用

作者姓名: 罗天宇

指导教师: 王石 正高级工程师

中国科学院计算技术研究所

学位类别: 电子信息硕士

学科专业: 计算机技术

培养单位: 中国科学院计算技术研究所

2026 年 6 月

Research and Applications of Causal Reasoning Based on Graph
Completion and Process Verification

A thesis submitted to
University of Chinese Academy of Sciences
in partial fulfillment of the requirement
for the degree of
Master of Electronic and Information Engineering
in Computer Technology

By
Tianyu Luo
Supervisor: Professor Wang Shi

Institute of Computing Technology, Chinese Academy of Sciences

June, 2026

中国科学院大学 学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。承诺除文中已经注明引用的内容外，本论文不包含任何其他个人或集体享有著作权的研究成果，未在以往任何学位申请中全部或部分提交。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日 期：

中国科学院大学 学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关收集、保存和使用学位论文的规定，即中国科学院大学有权按照学术研究公开原则和保护知识产权的原则，保留并向国家指定或中国科学院指定机构送交学位论文的电子版和印刷版文件，且电子版与印刷版内容应完全相同，允许该论文被检索、查阅和借阅，公布本学位论文的全部或部分内容，可以采用扫描、影印、缩印等复制手段以及其他法律许可的方式保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：

日 期：

导师签名：

日 期：

摘要

因果推理是理解世界运行规律、解释现象生成机制和支撑理性决策的重要能力。因果推理关注现象背后的生成机制、事件之间的作用关系以及条件变化下的结果演化，能够为复杂系统理解与决策支持提供机制性依据。从因果阶梯视角看，因果推理通常涉及观测关联、干预和反事实推演等不同层次的能力。

近年来，大语言模型（Large Language Models, LLMs）在自然语言理解、内容生成和复杂任务处理方面取得了显著进展，并逐步进入医疗、金融、信息治理和工业制造等高价应用场。如何以大语言模型为技术载体构建更稳定、可解释的因果推理方法，已成为当前研究中的重要问题。现有大语言模型的能力基础仍主要建立在统计相关性和语言模式匹配之上，难以充分理解因果关系的形成机制，也难以准确推断原因变化对结果的影响。现有方法在复杂因果推理任务中，仍存在因果机制建模不足和反事实因果推理过程失稳等问题。

围绕上述问题，本文开展了基于图谱补全和过程验证的因果推理研究与应用。因果推理包含因果判断、反事实因果推理等不同层次的任务：其中，因果判断主要关注因果对的识别与解释，反事实因果推理则是在既有因果对基础上，识别可能影响因果关系判断的潜在因素，并在这些因素保持不变的情况下，分析原因变化对结果的影响。基于这一认识，本文分别面向因果判断任务和反事实因果推理任务进行方法设计，并进一步在工业异常诊断场景中开展应用验证。本文的研究内容如下：

(1) 基于事理图谱补全的因果推理方法。针对大模型在因果判断中依赖统计相关性、易受表层线索干扰的问题，本文提出基于事理图谱补全的因果推理方法。该方法首先抽取并校验长文本中的事理图谱，在时序、条件、状态变化等非因果关系基础上补全事件间的潜在因果关系，通过强化关键因果路径和比对不同候选原因对结果的因果贡献生成因果摘要，从而将原始相关性判断转化为结构化因果比较。实验结果表明，该方法能够有效提升长文本情境下因果归因的准确性与稳定性。在三种参数规模的模型上，本文方法较最优方法因果判断正确率分别提升 13.5、11.9 和 8.9 个百分点。

(2) 基于过程验证的反事实因果推理方法。针对大模型在反事实因果推理中关键推理步骤缺失、条件约束信息遗漏的问题，本文提出基于过程验证的反事实因果推理方法。该方法围绕反事实推演中的关键中间过程，通过设计推理验证模块约束模型推理行为，引导模型按照完整推理链条完成分析。实验结果表明，该方法能够有效提升模型在反事实因果推理任务中的结果准确性与稳定性。在代码反事实任务和数学反事实任务上，本文方法较基线方法分别提升 3.83 和 6.56 个百分点。

(3) 基于因果推理的工业异常诊断系统框架。为验证前述方法在实际场景中的应用价值，本文面向工业生产异常诊断场景构建了基于因果推理的工业异常

诊断系统框架。基于事理图谱补全的因果推理方法主要用于异常成因识别、异常传播链分析和根因定位，为诊断过程提供结构化因果证据；基于过程验证的反事实因果推理方法主要用于根因识别后的条件变化分析，辅助推演关键工艺参数、设备状态或控制条件调整后可能产生的结果变化，从而支撑处置方案分析和改进建议生成。实证结果表明，上述方法能够在工业异常诊断链路中分别支撑原因识别、链路解释、条件推演和建议生成等环节，并为复杂工业场景下的智能诊断框架设计提供方法参考。

综上，本文围绕大语言模型在因果推理中的机制建模问题，提出了基于事理图谱补全的因果推理方法和基于过程验证的反事实因果推理方法。研究表明，本文提出的方法在长文本因果判断和反事实因果推理任务上具有较好的有效性，并在工业异常诊断场景中展现出较好的应用价值。本文的研究可为后续大语言模型因果推理方法研究及相关应用实践提供参考。

关键词：大语言模型，因果推理，图谱补全，过程验证，工业异常诊断

Abstract

Causal reasoning is an important ability for understanding how the world operates, explaining the generative mechanisms behind observed phenomena, and supporting rational decision-making. It focuses on the mechanisms underlying phenomena, the relations among events, and the evolution of outcomes under changing conditions, thereby providing mechanistic support for complex system understanding and decision support. From the perspective of the ladder of causation, causal reasoning usually involves different levels of capability, including observational association, intervention, and counterfactual reasoning.

In recent years, large language models (LLMs) have made remarkable progress in natural language understanding, content generation, and complex task processing, and have gradually entered high-value application scenarios such as healthcare, finance, information governance, and industrial manufacturing. How to construct more stable and interpretable causal reasoning methods with LLMs as the technical carrier has become an important research issue. However, the capability foundation of existing LLMs is still mainly built upon statistical correlations and language pattern matching, making it difficult for them to fully understand the formation mechanisms of causal relations and accurately infer the influence of cause changes on outcomes. Existing methods still suffer from insufficient causal mechanism modeling and unstable counterfactual causal reasoning processes in complex causal reasoning tasks.

To address the above issues, this thesis conducts research and application studies on causal reasoning based on graph completion and process verification. Causal reasoning involves different levels of tasks, including causal judgment and counterfactual causal reasoning. Among them, causal judgment mainly focuses on the identification and explanation of causal pairs, while counterfactual causal reasoning further identifies latent factors that may affect causal judgment based on existing causal pairs, and analyzes how changes in causes influence outcomes while keeping these factors unchanged. Based on this understanding, this thesis designs methods for causal judgment tasks and counterfactual causal reasoning tasks, respectively, and further validates them in industrial anomaly diagnosis scenarios. The main research contents of this thesis are as follows:

(1) A causal reasoning method based on eventic graph completion. To address the problem that LLMs tend to rely on statistical correlations and are easily disturbed by superficial cues in causal judgment, this thesis proposes a causal reasoning method based on eventic graph completion. The method first extracts and verifies eventic graphs from long texts, and then completes potential causal relations among events based on

non-causal relations such as temporal order, conditions, and state changes. By strengthening key causal paths and comparing the causal contributions of different candidate causes to the outcome, the method generates causal summaries, thereby transforming original correlation-based judgments into structured causal comparisons. Experimental results show that the proposed method can effectively improve the accuracy and stability of causal attribution in long-text scenarios. On models of three parameter scales, the proposed method improves the causal judgment accuracy by 13.5, 11.9, and 8.9 percentage points, respectively, compared with the best-performing baseline methods.

(2) A counterfactual causal reasoning method based on process verification.

To address the problems of missing key reasoning steps and omitted conditional constraint information in counterfactual causal reasoning with LLMs, this thesis proposes a counterfactual causal reasoning method based on process verification. The method focuses on key intermediate processes in counterfactual reasoning and constrains model reasoning behavior through a designed reasoning verification module, guiding the model to complete the analysis along a complete reasoning chain. Experimental results show that the proposed method can effectively improve the result accuracy and stability of models on counterfactual causal reasoning tasks. On code counterfactual tasks and mathematical counterfactual tasks, the proposed method improves performance by 3.83 and 6.56 percentage points, respectively, compared with the baseline methods.

(3) An industrial anomaly diagnosis system framework based on causal reasoning. To verify the application value of the above methods in practical scenarios, this thesis constructs an industrial anomaly diagnosis system framework based on causal reasoning for industrial production anomaly diagnosis scenarios. The causal reasoning method based on eventic graph completion is mainly used for anomaly cause identification, anomaly propagation chain analysis, and root cause localization, providing structured causal evidence for the diagnosis process. The counterfactual causal reasoning method based on process verification is mainly used for condition change analysis after root cause identification, assisting in inferring possible outcome changes after adjustments to key process parameters, equipment states, or control conditions, thereby supporting treatment plan analysis and improvement suggestion generation. Empirical results show that the proposed methods can respectively support cause identification, chain interpretation, condition inference, and suggestion generation in the industrial anomaly diagnosis pipeline, and provide methodological references for the design of intelligent diagnosis frameworks in complex industrial scenarios.

In summary, this thesis focuses on mechanism modeling in causal reasoning with LLMs, and proposes a causal reasoning method based on eventic graph completion and a counterfactual causal reasoning method based on process verification. The results show that the proposed methods achieve good effectiveness on complex text-based causal

judgment and counterfactual causal reasoning tasks, and demonstrate practical application value in industrial anomaly diagnosis scenarios. This study can provide a reference for subsequent research on causal reasoning methods for LLMs and related application practices.

Key Words: Large Language Models, Causal Reasoning, Eventic Graph Completion, Process Verification, Industrial Anomaly Diagnosis

目 录

第 1 章 绪论	1
1.1 研究背景	1
1.2 研究意义	2
1.3 研究问题	2
1.4 本文组织架构	3
第 2 章 相关工作 and 研究现状	5
2.1 引言	5
2.2 理论基础	6
2.2.1 Pearl 因果阶梯理论	6
2.2.2 “因果鸚鵡”假说	6
2.3 基于提示工程引导的因果推理方法	8
2.3.1 基础提示方法	9
2.3.2 思维链提示方法	10
2.3.3 因果特定的提示策略	12
2.4 基于模型微调的因果推理方法	14
2.4.1 因果知识与因果信号注入	14
2.4.2 显式因果结构监督	15
2.5 基于外部工具协同的因果推理方法	16
2.5.1 基于因果工具调用的增强方法	17
2.5.2 基于多智能体协作的因果推理方法	18
2.6 面向反事实推演的因果推理方法	19
2.7 大语言模型因果推理方法的总体进展	20
2.8 当前研究中的关键瓶颈	21
第 3 章 基于事理图谱补全的因果推理方法	23
3.1 引言	23
3.2 研究问题	23
3.3 相关工作	25
3.3.1 事理图谱	25
3.3.2 图谱补全	25
3.4 方法介绍	26
3.4.1 事理图谱抽取和校验	26

3.4.2 因果归纳边补全	28
3.4.3 因果摘要生成与使用	31
3.5 实验	32
3.5.1 数据集	32
3.5.2 模型构建和训练	33
3.5.3 实验结果与分析	34
3.6 小结	39
第 4 章 基于过程验证的反事实因果推理方法	41
4.1 引言	41
4.2 研究问题	42
4.3 相关工作	43
4.4 方法介绍	43
4.4.1 基于三阶段推理数据的监督微调方法	44
4.4.2 基于验证器信号的强化学习方法	45
4.4.3 过程级验证模块	46
4.5 实验	49
4.5.1 任务说明	50
4.5.2 数据构建	51
4.5.3 实验设置和训练超参数	54
4.5.4 实验结果与分析	55
4.6 本章小结	58
第 5 章 基于因果推理的工业异常诊断系统框架	59
5.1 应用背景	59
5.2 应用整体框架	59
5.2.1 整体业务流程	60
5.2.2 总体架构	60
5.3 应用案例分析	64
5.3.1 案例背景与问题描述	64
5.3.2 异常诊断过程分析	65
5.3.3 分析结果与应用效果	66
5.4 本章小结	67
第 6 章 总结与展望	69
6.1 本文工作总结	69
6.2 未来工作展望	70

参考文献·····	73
致谢·····	79
作者简历及攻读学位期间发表的学术论文与其他相关学术成果·	81

图目录

图 1-1	论文组织架构示意图	4
图 2-1	自回归生成顺序与因果顺序的错位 ^[1]	7
图 2-2	零样本提示和少样本提示 ^[2]	9
图 2-3	思维链方法 ^[3]	11
图 2-4	因果思维链 ^[4]	13
图 2-5	因果图微调 ^[5]	15
图 2-6	多智能体因果一致性框架	17
图 2-7	代码/数学反事实任务 ^[6]	19
图 3-1	基于事理图谱补全的因果推理方法	27
图 3-2	因果归纳边补全模型	28
图 3-3	事件摘要与因果摘要的消融结果比较	36
图 3-4	不同模型规模下各方法的相对增益	37
图 3-5	不同方法在细粒度因果因素上的响应	38
图 4-1	基于过程验证的反事实因果推理方法	44
图 4-2	归因验证器结构示意图	47
图 4-3	干预验证器结构示意图	49
图 4-4	代码反事实任务示例	50
图 4-5	监督微调与强化学习在三类任务上的总体表现	56
图 5-1	整体业务流程	60
图 5-2	总体架构	61
图 5-3	候选原因比较结果	66

表目录

表 2-1	因果阶梯	6
表 3-1	事理图谱关系类型	33
表 3-2	不同方法设置下的因果判断准确率比较	35
表 4-1	数学反事实任务示例	51
表 4-2	原始自由推理样本与三阶段推理样本示例	52
表 4-3	归因验证器标注样例	53
表 4-4	干预验证器标注样例	54
表 4-5	反事实任务上的实验结果	55
表 4-6	消融实验结果	57

表 5-1	溯因分析模块组成	62
表 5-2	异常任务背景信息	64
表 5-3	异常诊断任务包	65
表 5-4	异常诊断结果与处置建议	67

第1章 绪论

1.1 研究背景

因果推理是复杂系统理解和可靠决策支持中的重要能力。现实世界中的多数决策问题需要判断变量之间是否相关，还需要解释结果为何发生、识别关键影响因素、评估对原因进行干预可能带来的后果，分析原因变化时对结果的影响^[7-9]。因果推理关注现象之间的生成机制、作用方向与条件约束，是人类理解复杂系统和制定行动方案的重要认知基础，也是智能系统从相关性学习走向机制性理解的重要理论支撑^[10,11]。

因果推理的现实需求在高风险、强约束和多因素耦合场景中表现得尤为突出。在医疗领域，疾病发生、病程演化与治疗反应通常由病理机制、个体差异、治疗方案和环境因素共同作用决定，临床分析需要识别致病原因、评估治疗方法有效性、预测治疗方式的风险等^[12-16]。在金融领域，信用变化、风险传导、市场波动与欺诈行为具有复杂的链式影响和反馈效应，相关分析需要在预测结果之外进一步识别关键驱动因素与政策影响路径^[17,18]。在新闻传播、公共舆情和信息治理领域，信息扩散、叙事演化、平台推荐与群体行为之间存在显著的时序耦合与交互作用，治理过程需要理解传播机制^[19-21]。在工业制造领域，生产异常通常跨越设备、工艺、环境与操作等多个环节传播，异常定位、根因分析和参数调整同样依赖对作用机制的识别与推演^[22]。这些场景共同表明，因果推理对于支撑现实决策中的原因识别、机制解释和条件变化下的结果推演具有重要意义。

近年来，以大语言模型为代表的生成式人工智能快速发展，推动人工智能从面向特定任务的模型体系加速走向以基础模型为核心的通用能力底座^[2,23,24]。依托大规模预训练、指令对齐和持续扩展的工具协同能力，大语言模型在知识组织、自然语言理解、内容生成和复杂任务处理等方面展现出显著优势，并逐步进入教育、医疗、金融、政务、法律和工业等应用场景^[25,26]。随着大模型系统在现实业务中的参与程度不断提升，学术界与产业界对其推理深度、决策可靠性和应用安全性的关注持续增强。如何使大语言模型从基础自然语言处理进一步走向高价值场景分析与可信决策支持，已经成为人工智能研究发展的重要议题。

在因果推理研究中，大语言模型的发展带来了方法基础。现实场景中的因果信息往往分散存在于长文本叙述、领域文档、事件报告和交互问答之中，大语言模型在文本理解、语境建模、知识整合和语言交互方面的优势，使其有望成为承载复杂因果分析的重要技术载体^[27-31]。但是，现有研究指出，大语言模型在因果任务上的表现仍存在明显边界^[32-34]。模型在部分任务中能够给出表面合理的因果回答，但这种表现常常与内部知识调用、文本模式匹配和语料统计规律密切相关，其在复杂情境下的因果关系理解、使用干预推断因果关系、反事实推理等任务上的稳定性仍有不足^[4,35-38]。随着大语言模型不断进入高价值业务场景，以大语言模型为技术载体构建更稳定、可解释的因果推理方法，既有助于深化对

模型高阶推理边界的认识，也为推动因果推理方法走向复杂实际应用奠定基础。

1.2 研究意义

首先，研究以 LLM 为技术载体的因果推理方法可以拓展因果推理方法的适用领域。现实中的大量问题除了描述现象和预测结果，还要求系统解释结果为何发生、辨析哪些因素起到了关键作用，并进一步分析在条件发生变化之后结果将如何改变。从因果阶梯视角看，这类需求对应了从关联判断、到干预理解、再到反事实推理的递进过程^[7,39]。借助 LLM 强大的语义理解能力使得利用因果推理分析开放领域问题成为可能。

其次，结合因果推理可以推动大语言模型更加可靠、可信，并具有更好的可解释性。大语言模型具备较强的文本理解和语言生成能力，但其因果判断仍容易受到表层叙事顺序、局部显著线索和统计相关模式影响。围绕大语言模型的因果推理问题展开研究，有助于推动对模型能力的认识从表层语言匹配和统计共现利用，进一步转向对现象生成机制、原因结构、条件变化和结果推演的理解，也有助于深化对大语言模型推理边界的认识，提升模型在复杂推理问题中的分析深度和决策支撑价值^[32,34,35]。

此外，在实际场景中检验相关研究同样具有重要意义。实际业务场景中的因果分析问题通常具有证据来源分散、影响因素耦合、传播链路复杂等特点，更能够体现相关研究对复杂问题分析的支撑价值。将大语言模型相关研究进一步放到实际业务场景中加以考察，不仅有助于检验相关研究在现实环境中的可行性与适用性，也有助于推动大模型研究从理论分析走向工程实践。

1.3 研究问题

在上述背景与研究意义基础上，本文围绕长文本因果判断、反事实推理以及实际业务场景中的分析需求，提出以下三个研究问题：一是如何在长文本情境中形成较为稳定的因果判断，二是如何使 LLM 进行可靠的反事实因果推理，三是如何在实际业务场景中应用相关方法支撑复杂问题分析。三者共同构成本文的研究主线。

本文研究的第一个问题是：如何在长文本情境中形成较为稳定的因果判断。长文本情境中的因果判断任务要求模型从冗长文本中识别关键事件，理解事件之间的先后、依赖和竞争关系，并在多种候选解释之间形成较为稳定的归因判断。然而，长文本通常包含大量背景描述、时序信息、条件限制和竞争性原因，直接在原始文本空间中进行因果归因，容易受到表层叙事顺序、局部显著信息和冗余内容的干扰^[37,40-43]。针对这个问题，本文提出了基于事理图谱补全的因果推理方法，利用事理图谱抽取和事件间潜在因果关系的预测使得模型形成较为稳定的因果判断能力。

本文研究的第二个问题是：如何使 LLM 进行可靠的反事实因果推理。与一

般因果判断相比，反事实推理要求模型识别因果关系中未显式给出的隐藏变量，并在这些变量保持不变的前提下，准确推断原因变化对结果的影响。大语言模型处理这类问题时存在依赖文本表面相关性、常识联想和语料统计规律猜测答案等问题。针对这些问题，本文提出了基于过程验证的反事实因果推理方法，通过对 LLM 的推理过程进行关键步骤验证，从而提升其反事实推理的准确性和稳定性。

本文研究的第三个问题是：前述两项研究如何在实际业务场景中得到检验，并发挥对复杂问题分析的支撑作用。工业生产异常诊断任务具有跨工序传播、多源证据耦合、知识分布分散和分析目标明确等特点，相关任务不仅涉及异常成因识别，还要求对传播链路、关键因素及后续处置方案进行综合分析。这类场景能够更集中地体现因果判断与结果推演能力的实际支撑价值。本文将前述方法应用于工业生产异常诊断系统框架，验证方法落地的实用性。

1.4 本文组织架构

本文围绕基于图谱补全和过程验证的因果推理研究与应用展开论述，全文共分为六章。

第一章为绪论，主要对研究背景、研究意义、研究问题及全文组织结构进行总体说明。第二章为相关工作和研究现状，系统回顾因果推理的基本理论、大语言模型因果推理方法研究进展以及现有研究中的关键瓶颈，为后续研究提供理论基础与问题背景。第三章围绕长文本情境中的因果判断问题展开研究，提出基于事理图谱补全的因果推理方法，并验证其在长文本因果判断任务中的有效性。第四章围绕反事实因果推理问题展开研究，提出基于过程验证的反事实因果推理方法，并验证其在反事实推理任务中的有效性。第五章设计并实现了基于因果推理的工业异常诊断系统框架。将前面两个研究点的内容应用于工业异常诊断场景。并通过案例分析展示其应用价值。第六章为总结与展望，主要总结本文的研究内容，并对未来工作进行展望。

通过上述研究，本文形成了从问题提出、方法设计、实验验证到应用落地的完整研究链路，为面向大语言模型应用背景的因果推理方法研究及其工业场景应用提供系统性的研究基础。论文组织架构如图1-1所示。

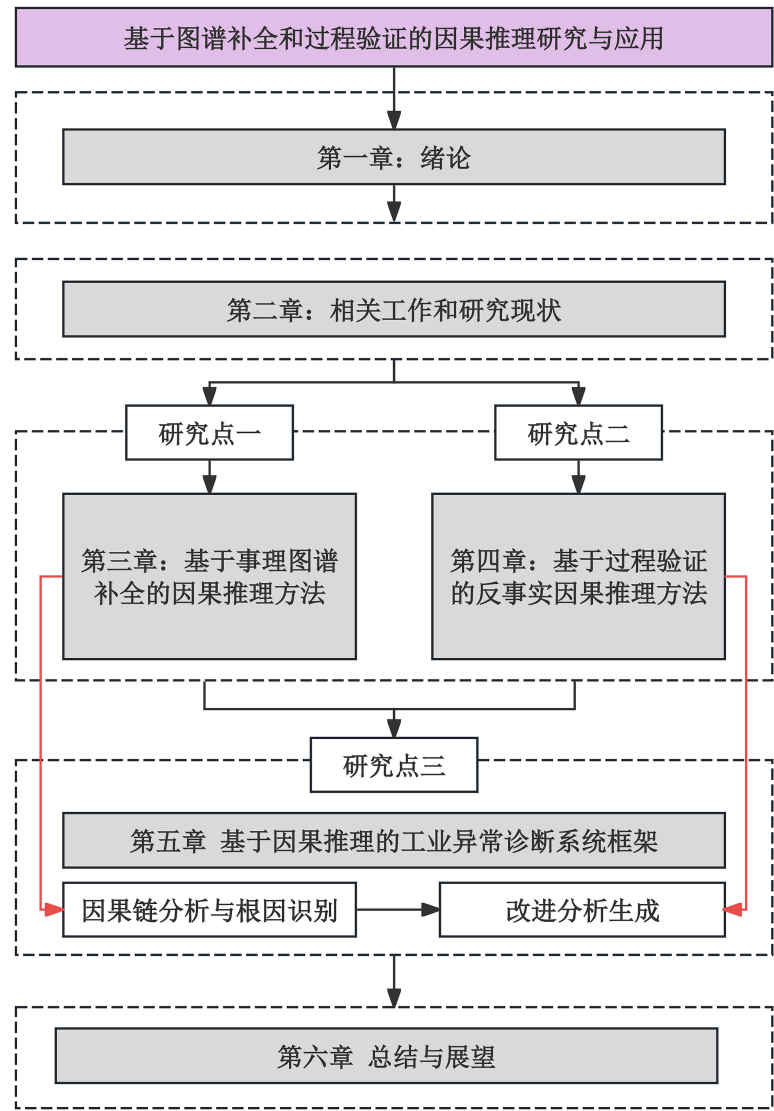


图 1-1 论文组织架构示意图
Figure 1-1 Thesis Organization Framework

第2章 相关工作和研究现状

2.1 引言

因果推理是人类理解世界运行规律的重要认知能力，对于揭示现象背后的生成机制、分析变量之间的因果关系以及评估干预效果具有重要意义^[7,10,11]。在人工智能研究中，赋予机器因果推理能力一直是推动智能系统迈向深层理解与自主决策的关键问题，同时也是一个具有显著挑战性的研究方向^[44,45]。近年来，大语言模型在自然语言理解、文本生成和复杂推理等任务中取得了显著进展，表现出强大的知识表征与任务泛化能力^[2,46,47]。随着相关研究的不断深入，LLM的因果推理能力边界，以及以大语言模型为技术载体的因果推理方法，逐渐成为学术界广泛关注的研究热点^[32-34]。

与一般的推理能力不同，因果推理具有独特的理论框架和方法论。Pearl提出的因果阶梯理论将因果推理分为三个层级：关联、干预和反事实^[9]。每个层级对应不同的认知能力和推理任务，从简单的统计相关性到复杂假设推理。这一理论为理解和评估LLM的因果能力提供了清晰的框架^[4]。

在大语言模型成为自然语言处理的重要技术载体之后，相关研究开始进一步考察模型能否真正处理上述不同层次的因果推理任务。然而，研究表明LLM在因果推理方面存在本质性的局限。Zečević等人提出了“因果鹦鹉”的概念，指出LLM虽然能够谈论因果关系，但其推理过程并非基于真正的因果理解，而是依赖于训练数据中因果事实的统计相关性^[35]。

LLM因果推理能力的不足严重制约了其在高风险领域的应用。在医疗诊断、政策制定和科学发现等场景中，准确的因果推理是做出可靠决策的基础^[12,15,17]。例如，当模型需要评估某种药物的疗效时，如果仅仅依赖于训练数据中的相关性而非真正的因果机制，可能会得出错误甚至有害的结论。因此，如何以大语言模型为主要技术载体，构建更稳定、可解释的因果推理方法，成为当前人工智能研究中的重要课题。

近年来，学术界围绕大语言模型背景下的因果推理方法开展了大量探索，形成了多样化的技术路线^[34,38]。这些方法可以分为提示工程引导、模型微调、外部工具协同和多智能体协作几类。提示工程方法通过精心设计的提示策略引导模型进行因果推理；模型微调方法利用因果知识、因果信号或显式因果结构约束模型训练过程；外部工具协同方法将因果发现、因果效应估计或形式化推断等工具引入模型推理过程；多智能体协作方法则通过角色分工、交叉审查和多轮讨论提升复杂因果分析的可靠性^[48]。不同方法各有优势和适用场景，但也面临各自的挑战。

本章将对当前大语言模型因果推理相关理论与方法研究进行回顾。首先介绍因果推理的理论基础与大语言模型因果能力边界，然后按不同技术路线对现

有大语言模型因果推理方法进行梳理和分析，最后总结当前研究中的关键瓶颈，为后续章节的研究提供理论基础与问题背景。

2.2 理论基础

2.2.1 Pearl 因果阶梯理论

因果推理与相关性分析存在本质区别。如表2-1所示，Pearl 等人提出的因果阶梯理论将因果关系划分为三个层级，每层对应不同的推理能力^[9]。

第一层级处于因果之梯的底层，其核心逻辑在于观测。该层级关注的问题可表述为：“若观察到变量 X 取特定值，变量 Y 的后验概率分布 $P(y|x)$ 是多少？”。处于关联层级的智能体仅处理变量间的统计共变性，无法区分伪相关与本质因果。现有的深度学习模型在此层级表现出色，能够从大量数据中学习复杂的关联模式^[49-51]。

第二层级为干预，涉及对系统的行动。对应问题“如果我操纵 X ， Y 会如何变化？”。数学上表示为 $P(Y|do(X))$ ，旨在量化对自变量 X 进行人为操纵后，因变量 Y 的响应情况，形式化表示为 $P(y|do(x), z)$ 。处于这一层级的智能体除了对数据的观察，还需理解变量间的因果方向和机制。例如，理解“如果强制给所有人接种疫苗，感染率会下降吗？”需要知道疫苗对免疫系统的因果效应，而非仅仅知道注射疫苗与感染率下降的相关性^[52,53]。

第三层级为反事实，涉及对已发生事实的想象。对应问题“如果 X 没有发生， Y 会怎样？”。这是最高层级的因果推理，要求在已知事实的基础上想象不同的可能性。数学表达为 $P(Y_x|x', y')$ ，即在观察到 $X = x'$ 和 $Y = y'$ 的条件下，计算在 $X = x$ 的干预下 Y 的取值。反事实推理是基于因果机制进行结果预测、因果归因和决策评估的重要基础^[39]。

2.2.2 “因果鸚鵡”假说

近年有相关研究提出了著名的“因果鸚鵡”假说，即大模型在因果问题上如同鸚鵡学舌，仅能复述训练数据中已存在的因果知识，而非真正理解因果机制^[35]。

研究指出 LLM 在某些情境下具备浅层的因果推理能力，这是因为 LLM 学

表 2-1 因果阶梯
Table 2-1 The Ladder of Causality

等级	公式	核心问题	典型活动	典型案例
关联	$P(y x)$	是什么？	观测	观测到发烧的人，患流感的概率是多少？
干预	$P(y do(x), z)$	如果... 会怎样？	行动	如果我强制给所有人接种疫苗，感染率会下降吗？
反事实	$P(y_x x', y')$	为什么？	想象	假如这位病人当初没有抽烟，他现在还会得肺癌吗？

习到了训练语料中已包含大量因果事实的文本表述。例如”海拔升高导致温度降低”这类经过人类因果归纳的知识。因此，LLM 在因果推理任务上的成功，源于其学习到了因果问题与答案之间的文本相关性，而非掌握了真正的因果机制。换言之，当 LLM 能够正确回答因果问题时，只是因为该因果事实恰好存在于训练数据中——它们”见过”这个答案，而非”推理”出了这个答案^[35,36]。

因果鹦鹉这一论断基于以下三个关键证据：

一、**自回归机制的非因果本质**。LLM 基于自回归生成机制，本质上是关联模型而非因果模型。具体而言，LLM 建模的是条件概率分布 $P(x_t|x_{s<t})$ ，这仅捕捉了变量间的统计关联。自回归机制的固有局限决定了 LLM 无法天然具备因果推理能力^[35]。

LLM 的语言建模过程如下：

$$\mathcal{L} = \sum_{t=1}^T \log P_{\theta}(x_t|x_1, x_2, \dots, x_{t-1}) \quad (2-1)$$

其中 x_t 表示序列中第 t 个词元， θ 为模型参数。该目标函数本质上刻画的是词元间的条件概率依赖，即在给定历史上下文时下一词元的统计分布规律，而非词元所承载语义概念之间的因果结构。

在模型的自回归训练中，序列当前值仅由过去值决定而与未来值无关，而通过自然语言表达的因果关系却未必满足此假设。从概率图视角审视，自回归模型隐含的图结构为有向链 $x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_T$ ，其中边 $x_i \rightarrow x_{i+1}$ 仅表示”生成顺序上的依赖”，而非”因果层面的影响”。

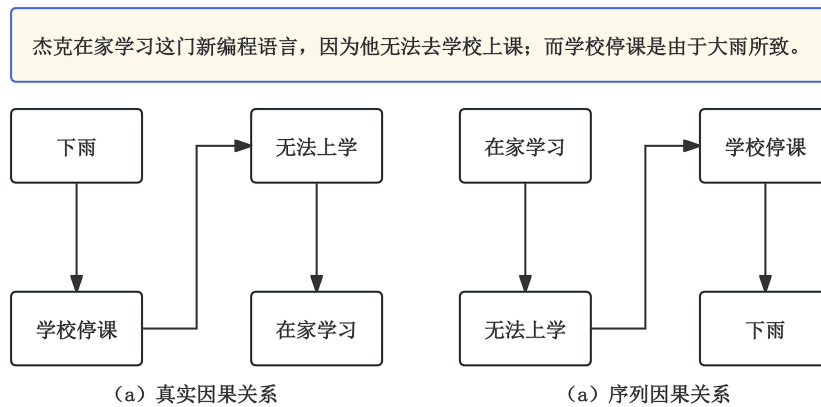


图 2-1 自回归生成顺序与因果顺序的错位^[1]

Figure 2-1 Misalignment Between Autoregressive Generation Order and Causal Order

如图2-1所示：”杰克在家学习这门新编程语言，因为他无法去学校上课；而学校停课是由于大雨所致。”该句涉及四个事件： $\{E_1$ ：下雨， E_2 ：学校停课， E_3 ：无法上学， E_4 ：在家学习 $\}$ 。因果图为 $E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow E_4$ 。然而，文本的词元

顺序为 $E_4 \rightarrow E_3 \rightarrow E_2 \rightarrow E_1$ ，与真实因果顺序完全相反。自回归模型在学习该句时，损失函数仅要求模型捕捉词元的共现统计规律，即 $P(E_1|E_2, E_3, E_4)$ 这类条件概率（序列因果关系），而不要求模型推断因果结构。

换言之，模型学会了“在怎样的上下文中出现怎样的词”，却无法回答“为何出现这些词”或“若改变某个事件，其他事件将如何变化”这类因果问题。

第二、LLM 因果能力本质。现有研究表明，LLM 在部分因果任务上的成功，往往更多依赖其对训练语料中因果表述模式的统计拟合，而非对因果机制本身的显式建模。Zečević 等人指出，模型在回答某些因果问题时，可能主要是在复述自然语言中已经存在的因果知识；Chi 等人进一步认为，当前 LLM 表现出的因果能力更多是浅层的、参数知识驱动的，而非稳健的人类式因果推理能力^[1]。

一个直观的说明性例子是：当被问及“海拔升高会导致什么”时，模型往往能够回答“温度降低”，但这种回答未必来自对大气物理机制的理解，更可能是因为训练语料中频繁出现了类似表述，模型据此学会了相应的文本映射关系。相反，当语料中存在误导性的相关模式时，模型也可能沿着这些线索生成错误的因果判断。例如，“巧克力消费量与诺贝尔奖得主数量”之间虽存在相关关系，但这种相关性可能同时受经济发展水平等混杂因素影响，因此并不能直接支持因果结论。若模型主要依赖表面统计模式，就可能难以识别这类伪相关。

第三、LLM 因果学习泛化性差。如果 LLM 真正掌握了可迁移的因果推理规则，那么当问题的表面形式发生变化时，其判断应当保持相对稳定。然而，现有研究表明，LLM 在因果任务上的表现仍然对变量命名、文本措辞和事件组织方式高度敏感^[36,37]。

Jin 等人在 Corr2Cause 基准上发现，尽管微调能够显著提升模型在分布内样本上的表现，但当研究者对问题中的变量名或文本表达方式进行扰动时，模型性能会明显下降。这表明模型并未稳定掌握“由相关推断因果”的抽象规则，而是较大幅度依赖训练中出现过的表层模式。类似地，有研究进一步发现，LLM 容易混淆时间关系与因果关系，且其判断结果会显著受到事件语言顺序的影响；即使语义关系不变，仅改变事件呈现方式，也可能引起模型输出波动。综合来看，当前 LLM 的因果推理能力仍具有较强的形式依赖性和分布敏感性：它们能够在熟悉表述下给出看似合理的答案，但一旦脱离训练中常见的表达模式，推理稳定性和泛化能力便明显不足。

然而，LLM 虽非“天生的因果模型”，却表现出极强的可塑性与情境学习能力。研究者们正致力于通过各种计算干预手段，试图在统计关联的底座上构建起具备逻辑一致性的推理表征^[34,38]。这一系列关于大模型因果推理能力的研究，下文将分别进行介绍。

2.3 基于提示工程引导的因果推理方法

提示工程方法无需修改模型参数，通过精心设计提示策略引导模型进行因果推理，具有实施成本低、适用范围广的优势。

2.3.1 基础提示方法

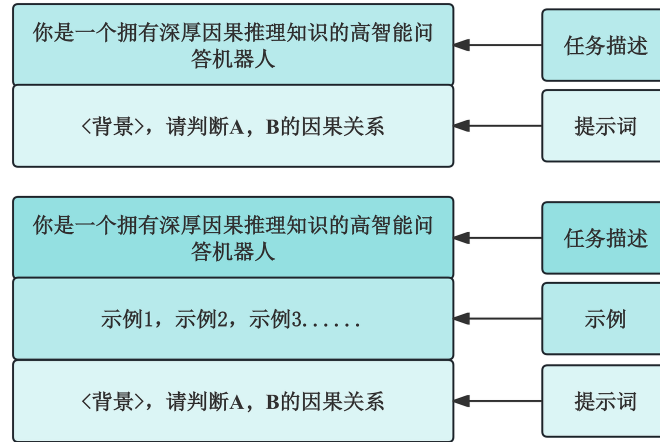


图 2-2 零样本提示和少样本提示^[2]

Figure 2-2 Zero-shot and few-shot prompting

如图2-2, 基础提示方法是指在不修改模型参数的前提下, 仅通过自然语言指令、少量示例或上下文组织方式引导模型完成特定任务的方法。相较于微调和工具集成方法, 这类方法具有实现成本低、部署灵活和迁移方便等优势。从提示设置上看, 基础提示方法主要包括零样本提示和 Few-shot 提示; 从作用机制上看, 这两类方法通常都可视为上下文学习范式在不同提示条件下的具体表现^[2]。

零样本提示是指在不给出任务示例的情况下, 直接以自然语言描述任务要求并要求模型作答。Brown 等人的研究系统展示了大规模语言模型在零样本和少样本提示设置下的任务适配能力, 表明模型能够在不进行参数更新的条件下, 仅依靠提示完成翻译、问答、完形填空等多种任务^[2]。在此基础上, 更多研究进一步表明, 在经过多任务指令微调后, 模型的零样本泛化能力可以得到明显提升^[24,25]。这说明零样本提示的有效性不仅取决于推理阶段的提示设计, 也与模型已有的指令遵循能力密切相关; 在经过指令微调的模型中, 这一点表现得尤为明显^[23]。

然而, 零样本提示的性能通常对提示措辞较为敏感。Reynolds^[54]指出, 精心设计的零样本提示在某些任务上甚至可能优于朴素的少样本提示, 提示本身并不仅仅承担任务说明的功能, 还会在一定程度上影响模型调用何种已有知识表征。对于因果任务而言, 这种提示敏感性表现得更为明显。Long^[55]等人的研究表明, 在因果图构建任务中, 仅改变提示前缀、关系动词或变量表述方式, 模型对变量间因果边存在性和方向性的判断就可能发生显著变化。Kıcıman^[56]等人也指出, 大语言模型在因果任务上的表现往往依赖具体提示形式, 且仍存在明显的不稳定性和不可预测的失败模式。因此, 零样本提示虽然能够快速激活模型中已有的因果知识, 但难以显式约束模型依据因果机制而非表层相关性作答。

Few-shot 提示是在提示中加入少量输入-输出示例的方式, 通常被视为上下

文学习最典型的实现形式之一。其核心作用在于利用少量示例向模型展示任务格式、标签空间和输出要求,从而帮助模型在当前上下文中完成任务适配。**Brown** 等人的工作表明,随着模型规模增大,少样本示例能够在部分任务上显著改善模型表现,因此 **Few-shot** 提示逐渐成为大语言模型任务适配的重要方式。与零样本提示相比, **Few-shot** 提示的优势在于能够通过示例为模型提供更明确的任务边界,降低输出形式的不确定性,并提高模型对当前任务模式的对齐能力。

不过,后续研究表明, **Few-shot** 提示的效果并不单纯取决于示例数量,还高度依赖示例本身的构造方式。**Min**^[57] 等人通过系统分析发现,在多种分类和选择任务中,示例中的真实标签并不总是决定模型性能的关键因素;相较之下,示例所提供的标签空间、输入分布和整体格式信息,往往对上下文学习效果更为重要。这说明 **Few-shot** 提示并不等同于传统意义上的小样本监督学习,其性能提升未必主要来自对少量样例中输入-输出映射关系的显式归纳,而可能更多来源于模型对任务形式和输出结构的快速识别。

除示例内容之外,示例选择策略也会显著影响 **Few-shot** 提示的效果。**Rubin**^[58] 等人指出,上下文学习的性能对所选示例高度敏感,并提出利用检索方法自动选择更合适的示例,以提升模型在语义解析任务上的表现。**Liu**^[59] 等人的研究进一步表明,与随机选取示例相比,基于语义相似度的示例检索往往能够构造出更有效的提示,从而在多种自然语言理解与生成任务上取得更好结果。由此可见, **Few-shot** 提示的有效性与示例是否具有代表性、相关性和合理的组织方式密切相关。

从作用机制上看,零样本提示与 **Few-shot** 提示均可视为大语言模型在上下文中完成任务适配的表现,这一过程通常被概括为上下文学习。其核心思想在于,模型无需通过梯度更新改变参数,而是在当前输入上下文中临时适应任务。**Brown** 等人的研究系统展示了大语言模型在该范式下的代表性能力,而 **Min** 等人、**Rubin** 等人和 **Liu** 等人的工作则进一步揭示了其作用机制与关键影响因素。总体来看,上下文学习的能力更多来自模型对任务格式、标签分布和输入模式的快速对齐。因此,该范式虽然具有较强的灵活性,但也容易受到提示形式、示例质量和任务表述方式变化的影响^[2]。

在因果推理场景中,基础提示方法表现出“易用但脆弱”的特点。一方面,这类方法可用于因果关系初步判断、因果信息抽取以及简单因果问答,能够以较低成本激活模型中已有的因果知识;另一方面,它们缺乏对混杂因素、干预机制和反事实结构的显式建模能力,因此难以保证模型稳定地依据因果机制而非表层相关性进行推断。特别是在更复杂的因果任务中,模型常常依赖语言表层模式而非稳定的因果规则作答^[60]。

2.3.2 思维链提示方法

思维链提示方法 (**Chain-of-Thought Prompting, CoT**) 的核心思想是在提示中显式引入中间推理步骤,将原本由问题直接映射到答案的过程转化为多步展开

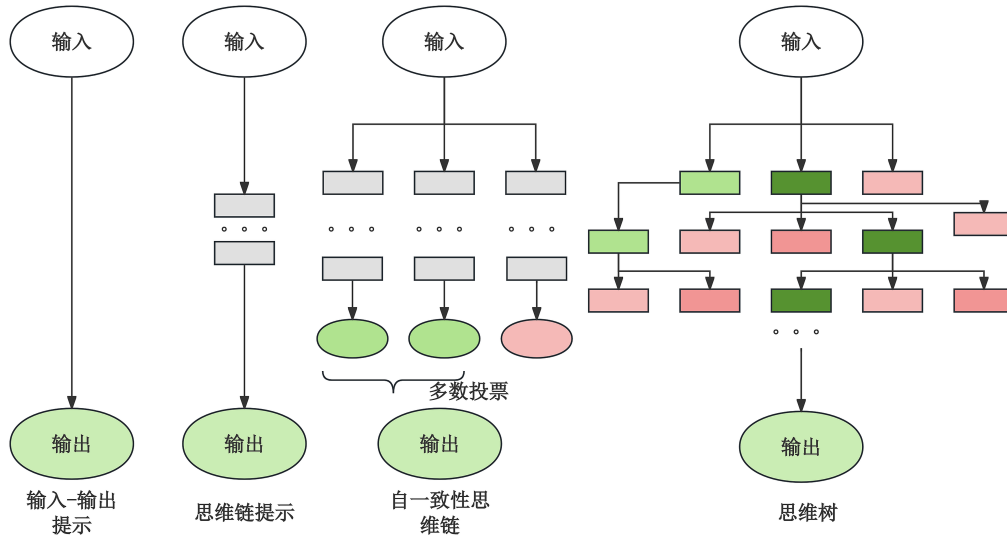
图 2-3 思维链方法^[3]

Figure 2-3 Chain-of-thought prompting

的求解过程。思维链提示通过显式生成中间推理步骤组织求解过程，其基本形式如图 2-3 所示。CoT 强调通过自然语言形式呈现中间推演路径，以降低复杂推理任务的求解难度。已有研究表明，在多步计算与复杂推理场景中，外显中间过程有助于模型完成较长推演链条。Wei 等人进一步将这一思想系统化为标准意义上的思维链提示范式，并将其界定为“一系列导向最终答案的中间自然语言推理步骤”^[61]。该方法通常在提示示例中同时给出问题、推理过程与最终答案，从而引导模型学习按照“分步推演—汇总结论”的方式完成任务。

从适用任务来看，CoT 主要用于算术推理、常识推理、符号推理等需要多步中间推演的场景。Wei 等人的实验表明，在足够大的语言模型上，引入思维链示例能够显著提升复杂推理任务的性能，其效果在数学应用题和符号推理任务上尤为明显^[47,61]。这说明 CoT 的关键作用是通过显式组织中间步骤，促使模型在生成答案之前先形成较为完整的求解路径。就方法本质而言，CoT 可被视为一种通过输出结构重组推理过程的提示策略，其优势在于将复杂任务拆分为若干相对可控的局部步骤，以缓解模型直接生成最终答案时可能出现的推理压缩问题。

在标准 few-shot CoT 基础上，Kojima 等人提出了零样本思维链方法，表明无需人工构造带有推理过程的示例，仅通过在问题后附加“Let’s think step by step”等触发语，也能够显著改善模型在多类推理基准上的表现^[62]。这一结果说明，大语言模型中可能已经存在可被提示激活的潜在多步推理能力，而 CoT 的作用之一正在于通过适当的语言信号将这种能力显化出来。不过，零样本思维链并不意味着模型已经稳定掌握了通用推理规则，其表现仍然会受到任务类型、模型能力及提示措辞等因素的影响。因此，更稳妥的理解是，零样本 CoT 揭示了模型在推理组织层面存在可激活的潜力，而非证明模型天然具备可靠且一致的深层推理能力。

随后,围绕 CoT 的改进研究主要沿着多路径聚合、问题分解与搜索式推理等方向展开。Wang 等人提出的自一致性方法以多条推理路径替代单一路径生成,通过对多个候选答案进行聚合来提高结果稳定性,其基本思想是:复杂问题往往存在多种可能的推理路径,而正确答案通常在这些路径之间具有更高的一致性^[63]。Zhou 等人提出的从简单到复杂提示方法,则将复杂问题拆解为一系列更易求解的子问题,并按照由易到难的顺序逐步完成推理,以缓解普通 CoT 在测试问题复杂度高于示例复杂度时的性能下降问题^[64]。Yao 等人进一步提出思维树方法,将线性的思维链扩展为可搜索的推理空间,使模型能够在不同候选“思维单元”之间进行探索、评估、前瞻与回溯,从而将推理过程由单路径生成推进到搜索式决策^[3]。针对零样本 CoT 中常见的漏步、计算错误和语义误解问题,Wang 等人又提出规划—求解方法,先生成整体计划,再依据计划逐步执行子任务,以提升零样本场景下思维链的完整性与可操作性^[65]。

总体来看,上述工作表明,思维链提示方法已经从最初的线性分步生成,逐渐扩展到多路径采样、结构化分解和搜索式推理等更丰富的推理组织形式。它们共同反映出一个基本趋势:CoT 的改进是更有效地组织中间状态、控制推理路径并提高最终答案的稳定性^[66]。

然而,CoT 并不等同于真实、稳定且完全可解释的推理过程。Paul^[67] 等人的研究表明,显式生成的思维链并不总是忠实反映模型内部的决策依据;在不同任务上,模型对推理链内容的依赖程度存在明显差异,有时会较强依赖中间步骤,有时则可能主要忽略推理链而直接生成答案。因此,CoT 虽然增强了推理过程的可见性,但其外显链条未必总能作为内部推理机制的可靠映射。除忠实性问题外,CoT 的效果还对提示示例与推理质量较为敏感。Chia^[68] 等人提出的对比思维链方法表明,在提示中同时展示合理与不合理的推理示例,有助于模型区分应遵循与应避免的推理路径,这从侧面说明示例质量与示例构造方式会显著影响 CoT 的表现。此外,由于 CoT 通常依赖自回归逐步生成,前面步骤中的错误还可能在后续推理中被持续放大,从而带来误差累积和额外计算开销^[69-71]。

尽管如此,思维链提示方法为模型提供了显式的中间求解结构,使复杂任务能够以更具层次性的方式展开,并为后续的多路径推理、规划推理和搜索式推理提供了统一的方法基础。对于本文关注的因果推理任务而言,CoT 所强调的“分步求解”思想同样具有重要启发意义,但因果推理不仅要求多步推演,还要求对因果图结构、干预关系和反事实语义进行明确建模,因此通用 CoT 往往难以直接满足形式化因果推断的要求。正因如此,研究者进一步提出了面向因果任务的专门提示策略,下一小节将对此进行讨论。

2.3.3 因果特定的提示策略

因果特定的提示策略,是指在提示设计中显式引入因果推断中的关键概念、结构要素或求解流程,引导模型更好地完成因果推理。这类方法更强调将混杂、中介、干预、反事实以及因果图等内容纳入推理过程,从而使模型输出不仅体现

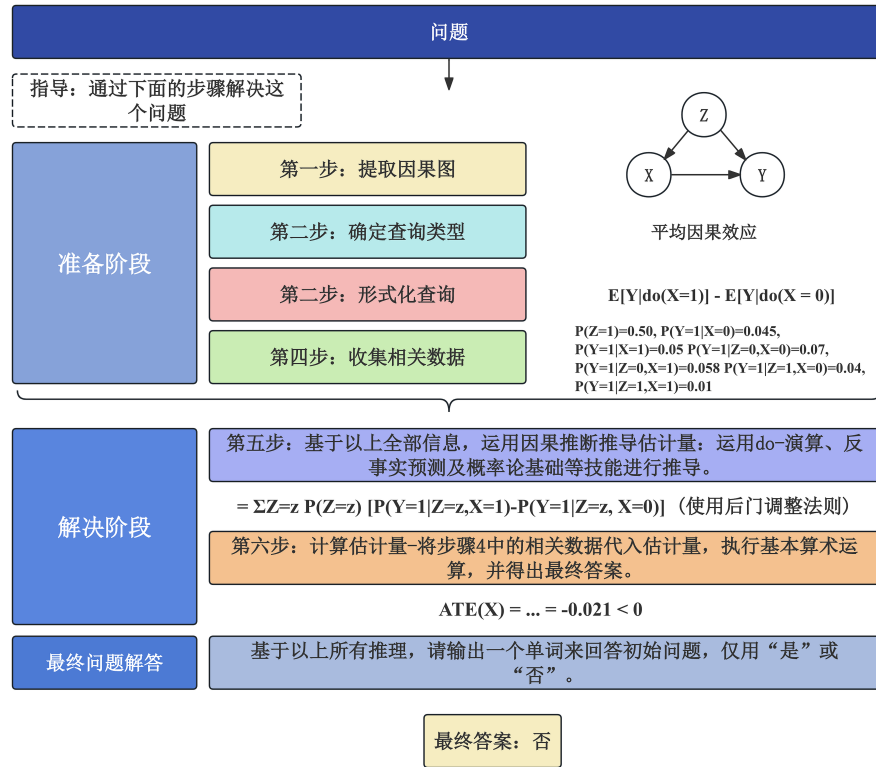
图 2-4 因果思维链^[4]

Figure 2-4 Causal chain-of-thought prompting

一般性的语言推演能力，也尽可能贴近因果问题本身的求解要求。当前较具代表性的思路，主要包括面向形式化因果查询的结构化提示策略，以及基于因果调整思想的提示去偏策略。

其中，CausalCoT 是该方向中与形式化因果推理结合最为紧密的代表性方法之一。该策略由 Jin 等人在 CLadder 的研究框架下提出，用于处理自然语言形式的因果推理任务，尤其关注关联、干预和反事实等不同类型的因果查询^[4]。其基本目标在于，将原本容易停留于表层语言相关性的问答过程，转化为更接近因果推断引擎的分步求解流程。CausalCoT 要求模型围绕因果图结构、查询类型以及可计算表达式开展组织化推理，从而为形式化因果问题提供更明确的任务约束。

CausalCoT 的求解过程可概括为准备与求解两个阶段。在准备阶段，模型首先需要识别问题中涉及的变量及其因果关系，并从自然语言描述中抽取潜在的因果图结构；随后判定当前问题属于关联、干预还是反事实查询，并进一步将原始自然语言问题转化为相应的形式化因果表达式；在此基础上，模型还需从题目中整理可用的概率信息、条件信息或干预信息，为后续计算建立输入条件。在求解阶段，模型依据前述形式化表示推导相应估计量，并结合给定数据或条件完成最终求值。由此可见，CausalCoT 的关键在于将因果推理中的结构识别、问题建模与估计量求解显式嵌入提示过程，使模型的中间步骤尽可能贴近因果推断的标准流程。

从方法作用看，CausalCoT 体现了通用思维链向因果任务专门化延伸的研究方向。它通过将自然语言问题分解为若干具有明确因果语义的中间步骤，增强了模型对因果任务求解结构的显式感知。尤其在形式化因果推理任务中，这种结构化提示有助于缓解模型直接从问题跳转到答案时所带来的推理压缩问题，并为分析模型究竟在哪一求解环节出现偏差提供更清晰的线索。然而，该方法的有效性仍高度依赖中间步骤的质量：若模型在因果图提取、查询类型判断或形式化转换阶段出现偏差，后续估计量推导与计算也会随之失真。因此，CausalCoT 虽然为因果任务中的结构化提示设计提供了较为清晰的实现框架，但尚不能保证模型稳定地执行忠实且正确的因果推断过程。

除 CausalCoT 外，另一类具有代表性的因果特定提示策略是 Zhang 等人提出的 Causal Prompting^[72]。该方法从因果调整视角重新审视提示过程本身。其基本思想是将提示、思维链与模型答案之间的关系置于结构因果模型中加以分析，并将思维链视为中介变量，尝试借助前门调整思想缓解潜在混杂因素对提示结果的影响。在实现上，该方法通常先生成多条思维链，再通过表示聚类选取代表性路径，结合检索得到的演示样本构造干预式提示，并对不同路径上的结果进行加权聚合。

总体而言，因果特定的提示策略体现了提示工程由一般语言引导走向因果结构引导的研究趋势。其中，CausalCoT 更强调按照因果推断流程组织问题求解，Causal Prompting 则更强调利用因果调整思想降低提示过程中的偏差。二者共同表明，在因果推理任务中可以借助显式的因果结构或因果假设来增强推理过程的可控性。不过，这类方法总体上仍属于推理阶段施加的外部约束，其效果容易受到任务形式、提示质量以及中间表示稳定性的影响^[4,73]。

2.4 基于模型微调的因果推理方法

微调方法通过在因果相关数据上训练模型，显式地注入因果知识或约束模型的推理过程。相较于提示工程仅能在推理阶段施加影响，微调方法能够更深入地改变模型的参数空间，将因果知识内化到 LLM 内部，引导模型更好的进行因果推理。然而，微调方法也面临数据标注成本高、因果知识难以形式化、以及在分布外场景泛化能力不足等挑战。现有工作主要沿两条路径展开：一是将因果知识注入预训练模型，增强模型对因果关系的感知能力；二是利用因果框架指导训练过程，帮助模型学习任务所需的因果表示^[73-77]。

2.4.1 因果知识与因果信号注入

因果知识与因果信号注入，是指通过预训练任务设计、监督目标构造或内部表示约束等方式，将与因果关系相关的外显知识或细粒度信号纳入模型训练过程，从而增强模型对因果线索的识别能力。这一路径更强调通过参数更新改变模型的表示偏好，使其在面对输入时更倾向于关注具有因果意义的信息。

较早的代表性工作 Li 等人提出的 CausalBERT。该方法注意到，传统预训

练语言模型虽然能够从大规模语料中学习丰富的语义表示，但其参数空间中并未显式编码稳定的因果关系结构。为此，作者从不同来源收集多粒度因果对，并围绕因果对分类与排序构造辅助训练目标，引导模型在参数层面学习“原因—结果”之间更具区分性的语义关联。实验结果表明，经过因果知识注入后，模型在因果对分类和因果问答等任务上的表现优于未注入相关知识的基线模型。这一结果说明，通过专门设计的训练任务将外显因果知识转化为模型内部表示，是增强语言模型因果敏感性的一条可行路径^[74]。

在此基础上，后续研究进一步将关注点由因果知识本身扩展到模型内部信息分配机制。Han 等人提出的因果注意力调优方法^[78]认为，大语言模型在处理复杂输入时，往往难以稳定区分真正具有因果解释力的词项与仅体现统计相关性的干扰线索，从而在分布变化条件下出现推断偏移。针对这一问题，该方法借助辅助模型自动构造 token 级因果监督信号，并通过 Re-Attention 机制对注意力分布施加约束，使模型在训练过程中更集中地关注因果相关词。与早期因果知识注入方法相比，这类方法进一步关注“模型在内部如何分配对因果信息的注意”。

总体来看，因果知识与因果信号注入方法主要从表示学习角度提升模型对因果关系的敏感性。其共同特征在于，不直接将形式化因果推断流程完整写入模型，而是先通过知识注入、监督约束和内部机制调整，提高模型对原因、结果及其关联线索的辨识能力。此类方法的优势在于较易与现有预训练模型结合，能够以较低的结构改造成本增强模型的因果表征；但其能力边界也较为明确，即更多停留在因果关系识别与因果线索筛选层面，对于变量间复杂依赖结构、干预语义和反事实条件的表达仍然有限。

2.4.2 显式因果结构监督

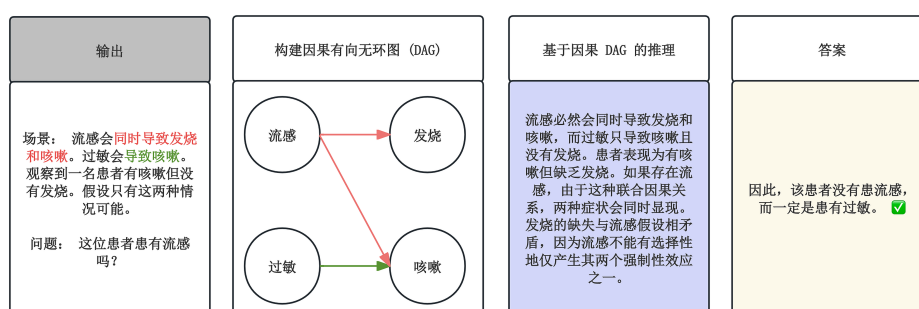


图 2-5 因果图微调^[5]

Figure 2-5 Fine-tuning with Causal-Graph Guidance

与因果知识和因果信号注入方法相比，显式因果结构监督更进一步地要求模型学习变量之间的结构关系，并在此基础上完成推理。前一类方法的重点在于增强模型对因果词项、因果对或因果证据的感知能力，而这一类方法则试图将因果图、变量依赖关系以及图上的求解过程共同纳入训练目标，使模型不仅能够识

别“哪些内容可能具有因果意义”，还能够进一步表示“这些变量如何构成因果结构”以及“应当依据何种结构化流程完成求解”。

这一研究路线的重要动因在于，通用语言推理方法通常主要在 token 序列层面展开。无论是直接答案生成，还是链式思维、思维树等推理组织方式，其核心操作对象都仍然是语言片段及其生成顺序。虽然这类方法能够在一定程度上改善多步推理表现，但对于变量之间的因果依赖关系、条件独立性、干预操作以及反事实条件的表达仍然缺乏足够明确的结构支撑。尤其在形式化因果推断任务中，若模型不能显式区分变量、边及其作用方向，就难以稳定地完成从自然语言问题到因果查询表示，再到结构化求解的全过程。因此，研究者开始尝试将因果图构建与图上推理作为监督目标的一部分，使模型从单纯的语言层推演转向变量层与结构层上的因果建模。

Li 等人提出的因果 DAG 构建与推理框架是这一方向中的代表性工作。如图2-5所示，该方法采用监督微调方式，训练模型首先从问题中识别因果变量及其关系，并进一步构建有向无环图；在此基础上，模型再依据所生成的因果结构执行后续推理。为支撑这一训练过程，作者构建了包含数万样本的 CausalDR 数据集，使每个样本不仅包含原始问题和答案，还包含相应的因果图及图上推理轨迹。与仅监督最终答案的方法相比，这种训练方式将变量识别、结构构建与求解步骤共同纳入学习目标，从而使模型能够在中间表示层面获得更明确的因果约束。实验结果表明，显式因果结构建模不仅提升了因果推理任务的性能，还在一定程度上缓解了模型输出中的逻辑不一致性幻觉问题^[5]。

总体而言，基于微调的因果增强方法试图通过参数更新将因果知识、因果信号或因果结构更深入地纳入模型内部表示，相较于仅在推理阶段施加约束的提示工程，这一路径在能力塑造上具有更强的持久性。无论是因果知识与因果信号注入，还是显式因果结构监督，其共同目标都在于减弱模型对表面相关模式的依赖，并提升其对因果关系及其求解过程的适配能力。不过，这类方法目前仍受到明显约束，其中最突出的瓶颈在于高质量因果数据与结构标注数据的获取成本较高。无论是因果对、token 级因果监督信号，还是变量关系、因果图结构与图上推理轨迹，其构建过程通常都需要较强的领域知识、形式化约束和人工校验，因而难以像一般指令数据那样实现大规模低成本扩展。与此同时，已注入的因果知识和结构模式往往仍与具体任务和领域场景密切相关，在分布变化条件下的泛化能力尚不稳定。正因如此，仅依赖参数更新仍难以完全弥合语言建模与形式化因果推断之间的差距，研究者进一步尝试借助模型外部的工具和协作机制，对因果求解过程进行补充与约束。

2.5 基于外部工具协同的因果推理方法

基于外部工具协同的因果推理方法，对大语言模型的因果求解过程进行补充、约束和验证。这类方法的重点在于，将外部模块纳入求解流程，使模型能够在外部支持下完成更具结构性的因果任务。现有研究主要包括两类路径：一类是

通过调用因果工具提升模型在结构化计算和因果操作方面的能力，另一类是通过多智能体协作强化因果推理过程中的审查与修正。^[26,48,79,80]

2.5.1 基于因果工具调用的增强方法

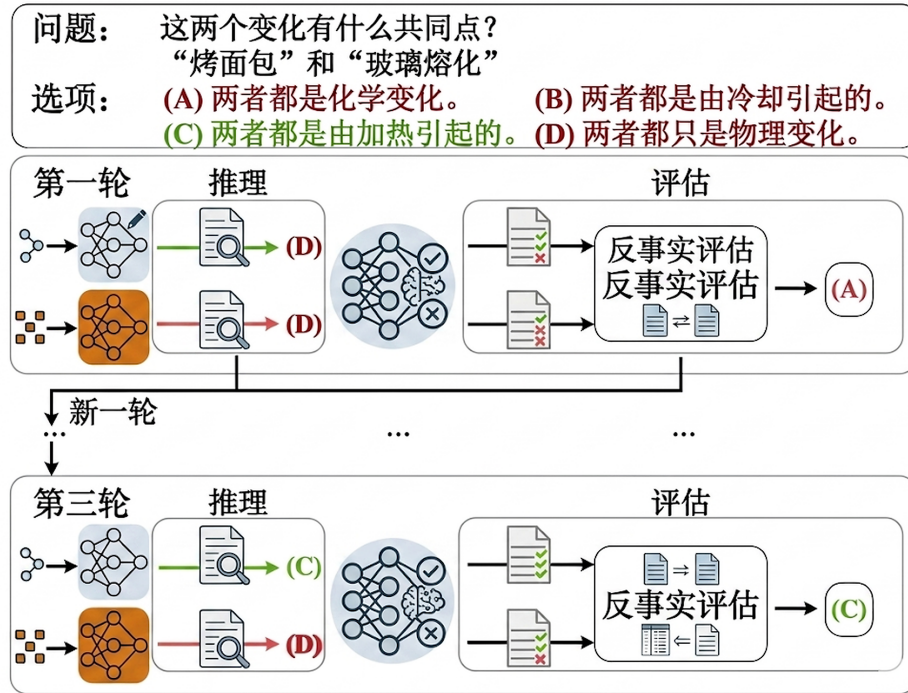


图 2-6 多智能体因果一致性框架

Figure 2-6 Causal Consistency Framework with Multi-Agent Collaboratio

基于外部工具协同的因果推理方法，是指在大语言模型求解因果任务的过程中，引入外部可执行因果模块、代码环境或结构化存储单元，以补足模型在因果图操作、表格数据处理和效应估计等方面的能力缺口。这类方法所接入的外部模块直接服务于因果问题的结构化求解。其基本思想在于，将原本难以仅凭自然语言稳定表达和执行的因果操作，转化为可调用、可计算和可验证的外部过程，从而使模型在面对变量级、边级、图级乃至因果效应级任务时，能够在系统支持下完成更可靠的推断。

这一研究路线的重要动因在于，大语言模型虽然在自然语言理解与生成方面具有显著优势，但因果问题往往同时涉及变量关系表示、图结构操作以及统计估计等内容，因而与语言模型擅长的文本生成任务之间存在明显差异。尤其是在因果图构建、因果效应估计和结构化因果问答等任务中，若模型仅依赖内部语言模式直接生成答案，往往难以稳定处理变量之间的形式化关系，也难以保证中间计算过程的可追踪性。

Han^[48] 等人提出的 Causal Agent 是这一方向中的代表性工作。该方法针对因果问题难以完全以自然语言表述、因果数据多为表格形式以及因果工具不易

直接嵌入语言模型推理流程等问题，在智能体框架中为模型配置了工具模块、记忆模块和推理模块。其中，工具模块通过 Python 调用和封装的因果函数实现具体因果操作，记忆模块以因果图为核心组织中间状态，推理模块则负责围绕任务需求协调各模块完成多轮求解。在这一框架下，大语言模型通过外部因果工具补足其在结构化求解上的不足。实验结果表明，该方法在变量级、边级、因果图级和因果效应级任务上均取得了较好的效果，说明因果工具调用能够在一定程度上缓解语言模型面对正式因果任务时的能力边界^[48]。

总体来看，基于外部工具协同的因果推理方法将因果图操作、效应计算和结构化推理等关键步骤外化为可执行过程，从而提高因果任务求解的可控性、可验证性与可追踪性。与提示工程和微调方法相比，这一路径通过外部工具补偿其在形式化因果操作方面的不足。不过，这类方法目前整体上仍处于较早阶段，相关研究数量相对有限，且系统性能在较大程度上依赖外部工具接口设计、模块协同效率以及中间结果表示的一致性。

2.5.2 基于多智能体协作的因果推理方法

已有研究表明，角色扮演、流程编排、辩论与动态团队组织等机制能够在一般复杂任务中提升协作效率与推理质量。Li 等人提出的 CAMEL^[81] 展示了基于角色扮演的自主协作框架，Hong 等人提出的 MetaGPT^[82] 则进一步强化了标准化流程和角色分工在复杂任务组织中的作用。Du 等人提出的多智能体辩论方法说明，多轮观点碰撞与修正有助于提升推理结果的稳定性；而 Liu 等人的 DyLAN^[83] 则从团队组织结构出发，尝试通过动态选择智能体提升协作效率。这些工作为后续因果推理方法研究提供了可借鉴的组织形式和交互机制，但其研究重点主要在于提升一般任务中的协作与推理表现，尚未围绕干预判断、反事实分析或因果一致性检验形成专门的协作范式。

在此基础上，Tang 等人提出的 Towards CausalGPT 则将因果一致性明确引入多智能体协作过程，是该方向更具代表性的工作。多智能体协作下的因果一致性推理流程如图 2-6 所示。该方法的基本出发点在于，知识密集型推理任务中的错误不仅表现为事实召回不足，还常表现为推理链中的逻辑跳跃、因果倒置和证据偏置，而传统单模型思维链方法通常缺乏对这些错误的外部审查机制。为此，作者提出因果一致性思维链框架 CaCo-CoT，通过多智能体协作对推理链进行生成、评估与修正，使模型在协作过程中接受针对因果合理性的额外检验^[80]。

CaCo-CoT 的核心结构由推理者与评估者两类角色构成。推理者负责围绕问题生成包含概念解释、子问题分解和逐步求解过程的结构化推理链；评估者则对候选推理链进行外部审查，其中既包括从非因果角度检验各步骤前提是否充分，也包括从反事实角度考察当候选答案发生变化时，既有推理是否仍然成立。前者主要用于识别条件不足、论证跳跃和因果方向混淆等问题，后者则进一步利用反事实视角检验证据对结论的支撑是否具有更强的因果解释力。通过这种“生成—审查—修正”的协作机制，因果一致性被嵌入推理链评估过程中。

不过，这类方法同样面临若干限制，例如多轮协作带来的较高计算开销、共享基础模型可能导致的系统性偏差，以及因果一致性评估本身对任务表述和语义差异较为敏感。

2.6 面向反事实推演的因果推理方法

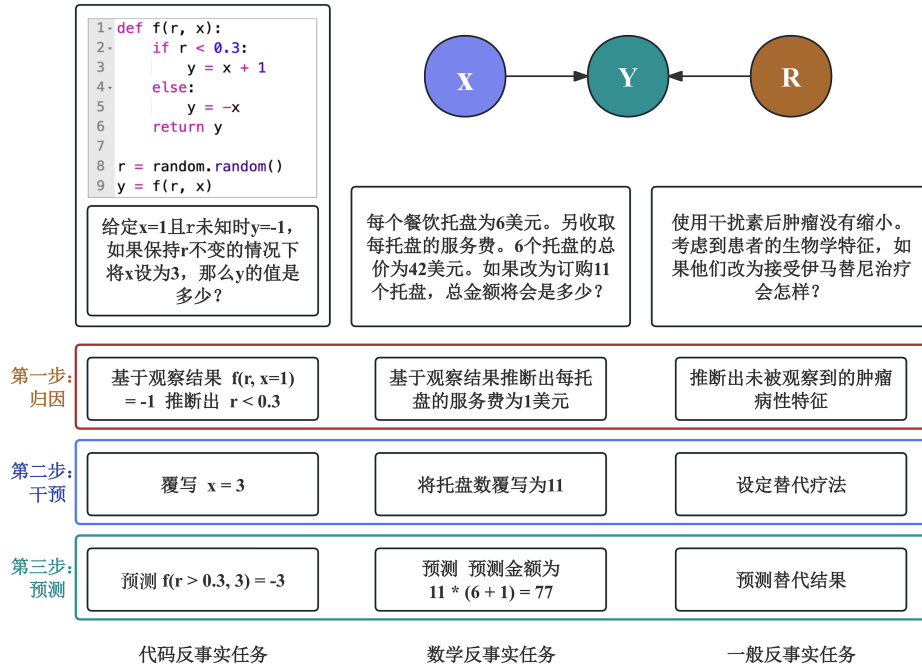


图 2-7 代码/数学反事实任务^[6]

Figure 2-7 Executable Counterfactuals

反事实因果推理是因果阶梯中的高层次能力，要求模型在既有事实成立的前提下分析条件变化后的可能结果。相比于常识性因果判断和相关性层面的因果问答，反事实因果推理不仅要求模型理解事件或变量之间的因果联系，还要求模型区分现实事实与替代情境，在给定假设条件下完成结果推导，并保持推理过程的一致性。因此，如何评估和提升大语言模型在反事实场景中的稳定推演能力，逐渐成为因果推理研究中的重要方向。

现有研究首先从任务构造和能力评测角度揭示了大语言模型在反事实推理中的能力边界。Frohberg 等人提出 CRASS 数据集^[84]，利用问答化的反事实条件句评估大语言模型的反事实条件理解能力，并通过人类基线进行对照，实验表明该任务对当时的主流模型仍构成明显挑战。Qin 等人提出反事实故事推理与生成任务^[85]，通过改变叙事中的关键事件，要求模型对原故事进行最小修改，使后续情节与给定反事实事件保持一致，从而考察模型对叙事因果链和反事实不变量的理解能力。Yu 等人构建 IFQA 数据集^[86]，将反事实预设引入开放域问答，要求模型在给定假设条件成立的前提下回答问题；这类任务要求模型不仅检索

相关知识，还要在可能违背已有参数知识的想象情境中完成推理。

随后，相关评测进一步从假设情境理解、逻辑修改和形式化因果查询等角度展开。Li 等人利用反事实条件句测试语言模型对假设场景的理解能力^[87]，指出这类任务能够区分模型是依赖统计相关和词汇线索，还是能够基于给定假设进行系统性推理。Huang 等人提出 CLOMO 任务^[88]，通过对论证文本中的逻辑条件进行反事实修改，考察模型在条件变化后维持逻辑关系的能力。Cladder 则将关联、干预和反事实查询纳入统一评测框架，使用形式化因果规则生成答案，用于评估模型在因果阶梯不同层级上的表现。进一步地，CounterBench^[89] 基于结构因果模型构造不同难度、不同图结构和不同问题类型的反事实查询，实验结果表明反事实推理对大语言模型仍具有较高挑战性，许多模型在该类任务上的表现接近随机水平。总体来看，这些评测研究表明，反事实因果推理通常比常识性因果判断和相关性层面的因果问答具有更高复杂度；模型需要在假设条件下区分现实事实与替代情境，抑制参数知识或表层词汇线索的干扰，并在条件变化后维持一致性推导。

在反事实推理方法研究方面，现有研究进一步从事实问题与反事实问题之间的对应关系出发，探索通过反事实反馈构造训练信号。Huyuk 等人围绕事实回答与反事实回答之间的一致性设计训练反馈，通过构造事实-反事实配对样本，并结合反事实反馈信号，引导模型学习不同情境之间的对应关系，利用事实情境和替代情境之间的差异，为模型提供更有针对性的训练信号，从而提升模型在反事实问答中的表现^[76]。

另一类研究尝试将因果规则以显式形式纳入模型训练过程。Vashishtha 等人提出公理化训练方法，通过构造包含因果前提、假设和结论的规则样本，将因果传递性、图结构约束等因果规则写入训练数据，使模型能够从结构化样本中学习更抽象的因果推理模式。公理化训练能够为模型提供更明确的因果约束，有助于提升模型在结构变化、推理链延长和复杂因果关系条件下的泛化能力^[77]。

此外，也有研究通过可执行任务环境构造更受控的反事实推理训练与评测方法。Vashishtha 等人进一步提出可执行的反事实任务^[6]，如图2-7所示，这项研究将反事实问题组织为代码和数学任务，要求模型在给定事实输入、事实输出和反事实输入的条件下，推断隐藏变量并计算替代情境中的结果。这类任务具有明确的执行规则和可验证结果，能够较稳定地区分事实观测、条件变化和结果预测等环节，为反事实推理训练和评测提供了更受控的环境。

总体来看，现有研究已经从任务评测、反事实反馈、因果规则训练和可执行任务环境等角度探索了大语言模型反事实因果推理方法，为进一步研究反事实推演过程建模与过程约束提供了基础。

2.7 大语言模型因果推理方法的总体进展

从方法演化看，现有研究已经围绕大语言模型背景下的因果推理任务形成了多种技术路线。基于提示工程的方法主要通过任务说明、示例组织、思维链提

示以及因果特定提示策略,引导模型在推理阶段显式关注因果结构和中间步骤,其优势在于实现成本较低、适配灵活,但效果往往对问题表述和提示形式较为敏感。基于模型微调的方法则进一步通过因果知识注入、因果信号约束和显式结构监督,在训练阶段塑造模型对因果关系、因果结构和推理规则的表示能力,使其在特定因果任务中获得更稳定的适配效果。基于外部工具协同的方法通过引入因果发现、因果效应估计和形式化推断等外部模块,使模型能够在自然语言理解和结构化因果计算之间形成协同。基于多智能体协作的方法则利用角色分工、结果审查和互相辩论等机制,对模型输出进行多角度检验,从而提升因果分析过程的可解释性和稳健性。面向反事实推演的相关方法进一步围绕反事实任务构造、反事实反馈、因果规则训练和可执行任务环境展开,为模型处理条件变化下的结果推演提供了更明确的任务形式和训练基础。

总体而言,这些研究从输入引导、训练监督、工具协同、智能体协作和反事实过程建模等角度推动了大语言模型背景下的因果推理方法发展,也为本文进一步研究基于图谱补全和过程验证的因果推理方法提供了基础。

2.8 当前研究中的关键瓶颈

结合前述方法与已有评测结果可以看出,当前以大语言模型为技术载体的因果推理方法仍面临若干关键瓶颈,主要集中在输入表达敏感、因果与相关混淆以及反事实推理能力不足三个方面。

首先,输入表达敏感仍然是制约模型稳定性的突出问题。大量研究表明,模型在因果任务中的表现对问题措辞、上下文组织方式和输入形式较为敏感;即便问题所对应的因果结构保持不变,仅改变变量命名、描述顺序或提示前缀,也可能导致模型给出不同判断^[36,37,90]。这说明现有方法仍难以稳定抓住与因果求解真正相关的关键信息,模型判断在相当程度上仍受到表面语言形式的影响。

其次,因果与相关混淆表明模型在识别真正致因线索方面仍存在明显不足。现有模型往往能够复现语料中高频出现的因果表述,但当问题中同时包含背景信息、共现模式或潜在干扰线索时,模型仍容易将统计相关误判为因果关系,或将局部语义关联误当作稳定因果机制^[35,36]。这一问题在复杂问题和多事件场景中尤为突出,因为随着文本长度增长和事件关系增多,真正支撑因果判断的关键证据更容易被非关键内容掩盖,模型也更容易据此作出偏差判断^[41-43]。

最后,反事实推理能力不足仍然是当前研究面临的另一项核心挑战。与一般因果判断相比,反事实任务不仅要求模型识别原因与结果之间的关系,还要求其在保持事实基础一致的前提下处理条件替换、隐藏状态恢复和结果重构,并在多个可能情境之间维持结构一致性^[85,86,88]。现有研究虽然已通过因果知识注入、事实—反事实对微调等方式在一定程度上改善模型表现,但整体上看,模型在反事实问题上的推理过程仍不够稳定,说明其在条件变化下组织中间过程和完成一致性推演方面仍存在明显不足^[76,77]。

总体而言,现有研究已从输入设计、模型训练、外部工具协同和多智能体协

作等角度推动了以大语言模型为技术载体的因果推理方法发展，但在相关性干扰和反事实推理能力等方面仍存在较为明确的瓶颈。后续研究仍需围绕这些关键问题，进一步探索能够支撑稳定因果判断和可靠反事实推演的方法。

第3章 基于事理图谱补全的因果推理方法

3.1 引言

尽管大语言模型在事实记忆、文本生成与一般推理任务中展现出较强能力，但其训练目标本质上仍建立在下一词元预测之上，对语料中的统计相关关系具有更强的拟合能力，而对作用方向、约束条件与生成机制的显式建模能力仍然有限。在涉及结果成因分析的任务中，这一能力边界使模型较难形成稳定且可靠的因果判断。

长文本情境中的因果判断在许多现实问题中具有重要价值。例如，在法律案例分析中，责任认定通常依赖对行为顺序、规范条件和事件链条的综合考察；在公共政策评估中，政策效果的解释需要区分真正发挥作用的干预因素与同时出现的背景变化；在工业异常诊断、医疗事件分析和安全事故追溯中，结果成因往往涉及多个对象、状态和条件的共同作用。上述问题的共同特点在于，支撑判断的证据通常分散于较长叙事之中，相关对象、状态变化与条件信息之间存在复杂联系，因而对模型的综合判断能力提出了更高要求。

长文本情境中的因果判断通常伴随更分散的信息分布和更复杂的关系组织。时间顺序、条件约束、对象关系、状态变化和规则说明等内容往往出现在不同位置，并通过省略、指代或补充说明相互连接。在这种情况下，模型不仅需要识别文本中出现了哪些对象与变化，还需要在较长叙述中形成较为稳定的因果判断。若模型主要依据局部共现关系、表面显著信息或叙事顺序进行作答，则支撑因果判断的关键线索容易被冗余叙述和表层相关模式所掩盖，从而影响最终判断的稳定性。

基于上述背景，本章围绕长文本情境中的因果判断问题展开研究，提出一种基于事理图谱补全的因果推理方法。该方法通过对文本中的关键对象及其关系进行结构化组织，先构建事理图谱表示，再在结构空间中建模候选因素对结果对象的潜在因果贡献，并进一步对支撑因果判断的关键路径进行归纳与筛选。通过这一过程，分散在叙事中的时间、条件、状态变化、阻断关系和替代解释等线索能够被统一表达为可比较的结构化证据，从而为后续判断提供更稳定的依据。

本章其余内容安排如下：首先对本章所关注的任务进行研究问题说明；随后介绍知识图谱、事理图谱及图谱补全相关研究；在此基础上，详细说明本文提出的基于事理图谱补全的因果推理方法；最后通过实验验证所提出方法在长文本因果判断任务中的有效性。

3.2 研究问题

本章关注长文本情境中的因果判断任务。给定一段文本情境 C 以及文本中待判断的一对对象 x_a 与 x_b ，其中 x_a 表示候选原因对象， x_b 表示结果对象，模

型需要输出关于二者之间是否存在因果关系的判断结果 y 。该任务通常以问答形式呈现，可简要表示为：

$$(C, x_a, x_b) \rightarrow y \quad (3-1)$$

其中， y 表示模型对 x_a 与 x_b 之间因果关系的最终判断结果。下面给出一个长文本因果判断任务的示例。

例 3.1. 老板有一个很大的花园，并雇了员工甲和员工乙两名园丁照看植物。两人平时都很可靠，老板对他们的工作也一直很满意。工具棚里有几个常用喷壶，其中蓝色喷壶和绿色喷壶最常被拿来给花床施肥。按照平时分工，员工甲通常处理靠近温室的花床，员工乙通常处理靠近水池的花床。

春天刚开始时，老板对几种新肥料很感兴趣。他先拿肥料 B 在水池旁边最外侧的两排花上试了一次，之后把绿色喷壶清洗干净后放回工具棚。

几天后，老板在一本园艺杂志上看到，肥料 A 和肥料 B 都能促进植物生长，但如果同一片植物先后接触不同类型的化学肥料，可能会出现枯萎。读完这篇文章后，老板决定这次正式施肥只使用肥料 A。他告诉员工甲和员工乙：“这次花园里只用肥料 A，不要混用别的肥料。”员工甲说自己可以去购买肥料 A，并负责当天的施肥准备，包括把喷壶摆好、确认每个人使用的工具。按照花园里的准备流程，正式施肥前通常还要把要用的喷壶重新冲洗一遍。员工乙则负责水池一侧的花床。

几天后，员工甲买回了肥料 A。施肥当天早上，员工甲赶着在下雨前开始工作，只重新冲洗了自己要用的蓝色喷壶，忘了再冲洗绿色喷壶。他把蓝色喷壶和绿色喷壶都摆在工具棚门口，并对员工乙说：“水池那边还是用你平时那个。”随后，员工甲拿走蓝色喷壶去处理温室一侧的花床。员工乙看到自己平时使用的绿色喷壶也在门口，便拿着它去处理水池一侧的花床。

几周后，老板发现花园里的情况很不一致：有些植物长得更高、更漂亮，但另一些花床里的植物开始枯萎。

问题：员工甲的行为是否导致了植物的枯萎？

该示例用于展示长文本情境下因果判断任务的主要难点。文本中涉及老板、员工甲、员工乙、蓝色喷壶、绿色喷壶、肥料 A、肥料 B 以及不同花床区域等多个对象，并通过“你平时那个”“它”“水池那边”等表达形成跨句指代，模型需要持续追踪人物、工具、肥料和区域之间的对应关系。同时，关键因果线索分散在不同段落中：老板曾用肥料 B 处理水池旁边部分花床，杂志说明同一片植物先后接触不同化学肥料可能导致枯萎，正式施肥时员工乙又处理水池一侧花床，模型需要整合这些远距离线索，才能恢复“部分植物先前接触肥料 B，后来又接触肥料 A，因此发生枯萎”的因果链条。此外，员工甲“忘了再冲洗绿色喷壶”这一显著情节容易形成干扰，但文本已说明绿色喷壶此前被清洗干净，因此不能简单将枯萎归因于喷壶残留。

由此可见，长文本因果判断任务要求模型完成共指解析、长距离证据整合、干扰线索排除和候选原因比较等。现有大语言模型在处理这类任务时，往往更容易依赖文本中的表层相关模式、局部显著线索和叙事顺序进行判断。因而容易将与结果对象距离更近、表述更突出或共现更频繁的内容当作因果依据，从而影响最终判断的稳定性与准确性。

因此，本章的核心研究问题是：如何在长文本情境下减少模型对表层相关模式的依赖，使其能够更多依据文本中的结构关系、约束信息和传播线索形成更稳定的因果判断结果。围绕这一问题，本章进一步引入结构化建模思路，通过对文本中的关键对象及其关系进行显式组织，将原始文本中的开放式相关性判断转化为更可比较的因果判断问题。

3.3 相关工作

3.3.1 事理图谱

知识图谱通常以三元组形式显式表示对象及其关系，为非结构化文本提供结构化表达，便于后续检索、对比和推理。围绕文本到图谱的构建，早期研究主要依赖关系抽取和开放信息抽取等信息抽取方法，从句子中识别实体、关系及其组合，为知识图谱构建提供基础。随后，相关研究逐渐从句子级抽取扩展到文档级抽取与联合抽取，使模型能够在更长文本中同时建模实体、关系与事件信息，从而提升复杂文本场景下的结构化表示能力。

在此基础上，事理图谱进一步以事件作为核心节点，将事件之间的时序、条件、参与者关联及依赖关系显式组织起来^[91]。相关研究表明，事件中心化的表示方式能够更有效地刻画叙事文本中的结构信息，因此被广泛应用于事件模式归纳、叙事理解和复杂推理等任务。近年来，随着大语言模型的发展，研究者开始尝试利用提示生成、分步抽取和结果校验等方式，直接从原始文本生成事件三元组、事理图谱或事件模式，从而降低传统信息抽取流程的构建成本。

3.3.2 图谱补全

知识图谱补全旨在基于已有三元组推断缺失事实，是知识推理中的基础问题之一。早期方法主要基于知识图谱嵌入，通过低维向量表示实体与关系，并利用打分函数判断三元组是否成立。这类方法具有结构简单、训练效率较高等优点，因此成为知识图谱补全研究的重要基础。随后，研究者进一步引入图神经网络和路径推理机制，使模型能够利用邻域结构和多跳关系模式进行补全，从而增强对复杂关系模式的建模能力。

随着预训练语言模型和大语言模型的发展，知识图谱补全研究开始进一步结合文本语义信息。相关工作尝试利用实体描述、关系文本以及文本化的图事实来增强补全能力^[92,93]，从而缩小自然语言语义空间与图结构空间之间的差距。近年来，一些研究进一步强调结构感知建模，通过对齐文本表示与图结构表示，使语言模型能够在补全任务中更充分地利用结构信息^[94]。

3.4 方法介绍

大语言模型在长文本情境下的主要缺陷，表现为对分散信息、叙事顺序和复杂事件关系的利用不稳定。即使模型具备较长的上下文窗口，其对关键信息的提取仍会受到信息位置影响：当相关证据位于上下文中部时，模型更容易遗漏或弱化该信息，从而导致推理性能下降^[41]。在涉及事件关系的文本中，模型还容易受到语言呈现顺序的干扰，将事件的先后排列误认为因果方向，或者在时间关系与因果关系之间产生混淆^[37]。当文本进一步扩展为包含多个事件、跨句关系和反常识因果结构的叙事时，模型往往依赖事件出现顺序、参数化世界知识和局部启发式进行判断，难以稳定整合全文中的因果线索；叙事长度增加、事件数量增多以及文本因果结构与常识知识不一致，都会进一步放大这类推理失败^[95]。

针对大模型容易依赖表层相关性进行因果判断的问题，本文提出一种基于事理图谱补全的因果推理方法。该方法分为三个阶段：首先，从长文本中抽取只包含非因果关系的事理图谱，并通过一致性校验得到较稳定的图表示。这一步将原始长文本中的事件、条件、状态变化、时间先后和阻断等非因果关系显式组织为图结构，使分散在不同位置的关键线索能够在统一结构中被访问和比较，从而减弱模型对上下文位置和局部显著信息的依赖；然后，使用“因果归纳边补全模型”在事理图谱上估计候选事件对结果事件的潜在因果关系。该模型已经从训练数据中学习非因果关系与潜在因果关系之间的统计规律，能够对可能原因与结果之间的因果关联强度进行初步估计；最后，把补全得到的潜在因果关系和事理图谱中的事件间关系写入提示词，引导模型围绕候选原因与结果事件之间的关系进行细致的分析，并得到因果摘要。例如：说明高因果强度候选原因如何影响结果，对比不同候选原因并排除干扰事件等。这一步围绕统计上的因果可能性和事件间的叙事链路，通过对比、排除、链路完整性检查等方式凝练原本分散在长文本中的支持因果判断的线索。得到因果摘要后，本文将其与原始问题一起输入模型，帮助模型给出最终因果判断。

综上，方法第一阶段侧重从原始文本中抽取事件节点及其之间的非因果关系，第二阶段侧重在这些事件语义和结构信息基础上估计候选事件对结果事件的潜在因果关系，第三阶段侧重将潜在因果强度和事件链进一步组织为面向问题的整体证据，最后通过提示词的方式使用整体证据辅助模型进行因果判断。整体流程如图 3-1 所示。

3.4.1 事理图谱抽取和校验

这一阶段先将长叙事文本转化为可用于后续因果建模的事理图谱，再从多次抽取结果中选出一个代表性图谱。这一阶段的目的是收集散落在长文本中支持因果判断的关键线索，并通过图结构减少文本表述顺序对因果判断的影响。

为了服务后续因果判断，抽取出的事理图谱至少需要包含问题中的**候选原因事件**、**结果事件**以及与结果事件相关的**混淆事件集合**，即可能对结果形成替代解释的其他事件集合。其中，候选原因事件与混淆事件集合共同构成候选原因集

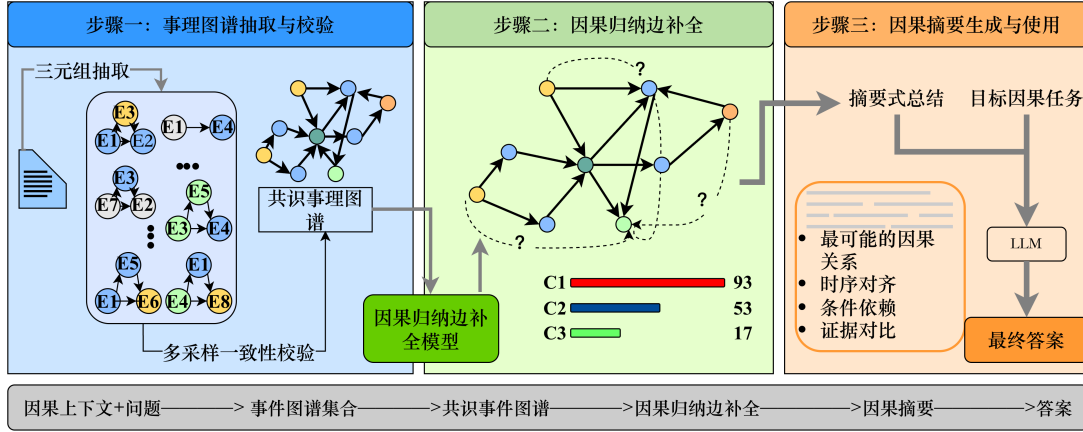


图 3-1 基于事理图谱补全的因果推理方法

Figure 3-1 Causal reasoning pipeline based on eventic graph completion

合，用于表示可能导致结果事件的不同解释路径。

抽取输出由三部分构成：

- (1) **候选原因事件与结果事件**：分别对应问题中的候选原因和待解释结果。
- (2) **混淆事件集合**：与结果事件高度相关、且可能形成替代解释的事件或条件集合，用于显式表示竞争性原因与背景因素。
- (3) **关系三元组**：以 [事件₁, 关系, 事件₂] 的形式表示事件之间的依赖关系，由此构成事件图 $G = (V, E)$ ，其中 V 表示事件节点集合， E 表示关系边集合。

在关系抽取阶段，本文方法要求模型优先抽取能够较完整表述事理图谱结构的关系类型，例如时间先后、条件触发、功能阻断、组成包含、状态变化和一般影响等。这样做是为了先把文本中的事件结构整理出来，而不在抽取阶段直接判断因果关系。因果判断放到后续的因果边补全过程中完成，可以减少抽取提示对最终判断的直接影响。

考虑到事理图谱由生成式模型抽取得到，同一输入在多次采样下可能出现事件表述波动、局部边差异或关系措辞变化。为减弱单次采样噪声对下游建模的影响，本文方法对同一输入文本进行 K 次独立采样，得到 K 个候选事件图，再从中选取一个代表性图谱。设第 k 次采样得到的关系三元组集合为 $\mathcal{T}^{(k)}$ ，则：

$$\mathcal{T}^{(k)} = \{(e_{h_m}, r_m, e_{t_m})\}_{m=1}^{M_k} \quad (3-2)$$

其中 e_{h_m} 与 e_{t_m} 分别表示第 m 个三元组的头事件和尾事件， r_m 表示对应的事件间关系，因此每个 $\mathcal{T}^{(k)}$ 都对应一个候选事件图。

为了比较不同候选图是否表达了相近的事件结构，本文先将每个候选图编码为向量，再计算候选图之间的语义相似度。具体地，首先对每个关系三元组构造语义表示。设 $\phi(\cdot)$ 为三元组编码函数，其输入为三元组 (e_i, r_{ij}, e_j) 的文本描述，输出为对应的向量表示 $\mathbf{t}_{ij}^{(k)}$ 。在实现上，可将头事件、关系和尾事件序列化为单段文本后输入预训练文本嵌入模型进行编码，即：

$$\mathbf{t}_{ij}^{(k)} = \phi(e_i, r_{ij}, e_j) \quad (3-3)$$

在得到三元组级表示后，进一步对候选图内全部三元组向量进行池化，得到图级表示：

$$\mathbf{g}^{(k)} = \frac{1}{|\mathcal{T}^{(k)}|} \sum_{(e_i, r_{ij}, e_j) \in \mathcal{T}^{(k)}} \mathbf{t}_{ij}^{(k)} \quad (3-4)$$

基于图级表示，任意两个候选图之间的相似性定义为余弦相似度：

$$s(\mathcal{T}^{(i)}, \mathcal{T}^{(j)}) = \cos(\mathbf{g}^{(i)}, \mathbf{g}^{(j)}) = \frac{(\mathbf{g}^{(i)})^\top \mathbf{g}^{(j)}}{\|\mathbf{g}^{(i)}\| \|\mathbf{g}^{(j)}\|} \quad (3-5)$$

这样可以从整体上比较不同候选图是否表达了相近的事件结构，并减弱事件表述和关系措辞波动带来的影响。在此基础上，进一步以候选图与其余所有候选图的平均相似度作为代表性分数：

$$\text{Score}(\mathcal{T}^{(k)}) = \frac{1}{K-1} \sum_{j \neq k} s(\mathcal{T}^{(k)}, \mathcal{T}^{(j)}) \quad (3-6)$$

最终，将代表性分数最高的候选图选为代表性事件图：

$$k^* = \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} \text{Score}(\mathcal{T}^{(k)}), \quad \tilde{\mathcal{T}} = \mathcal{T}^{(k^*)} \quad (3-7)$$

上述方法用于从多个候选图中选取代表性事理图谱。若某次抽取得到的候选图与其余样本保持较高平均相似度，则说明其更接近多次采样结果中的共同语义模式，因而更适合作为后续因果边补全模型的输入图谱。通过这一过程，本文从多次抽取结果中选出一个更稳定的事理图谱，作为后续模型的输入。

3.4.2 因果归纳边补全

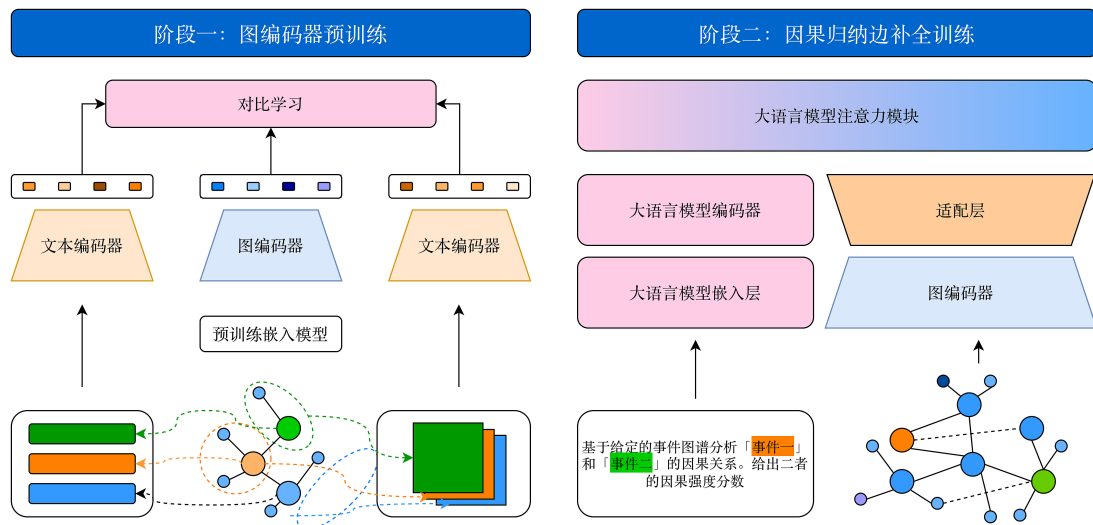


图 3-2 因果归纳边补全模型

Figure 3-2 Causal induction edge completion model

上一阶段得到的事理图谱主要包含时间先后、条件触发、阻断、状态变化等非因果关系。这一阶段进一步在这些非因果关系的基础上，补全事件对之间潜在的因果关系，尝试从相关性建模因果性，用于支撑最终因果判断。

本文将这种综合事件语义和非因果关系后得到的潜在因果关系记为**因果归纳边**。所谓因果归纳边，是指在事件图 $G = (V, E)$ 中，为事件对 (e_a, e_b) 建立的一条潜在因果关系估计边，用于吸收事件对之间存在的时序、组成、状态变化等非因果关系，以衡量事件 e_a 对事件 e_b 的可能因果贡献强度。该强度由带有因果强度标注的数据进行监督学习，并用于后续长文本因果判断任务中的预测。

本文参考 SAT 的两阶段框架^[94]，构建因果归纳边补全模型，用于预测事件图上事件对之间的因果归纳边强度。因果归纳边补全模型的整体结构如图 3-2 所示。模型训练分两个阶段进行：第一阶段学习事理图谱中事件和关系的表示；第二阶段利用这些表示预测给定事件对之间的因果归纳边强度。

第一阶段：图编码器训练。第一阶段的目标是学习更好的事理图谱节点表示。具体来说，先将事件节点和关系文本编码为向量，再通过图编码器聚合邻居节点和边关系的信息，得到每个节点的结构表示；随后利用对齐损失进行训练，使该表示同时保留节点语义和图结构信息。

设事件图中第 i 个节点对应的文本描述为 e_i ，边 (i, j) 上的关系文本为 r_{ij} 。记预训练文本编码器为 $\phi(\cdot)$ ，则节点与关系的初始语义向量分别为：

$$\mathbf{x}_i = \phi(e_i), \quad i = 1, 2, \dots, |V| \quad (3-8)$$

$$\mathbf{r}_{ij} = \phi(r_{ij}), \quad (i, j) \in E \quad (3-9)$$

将全部节点表示记为 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_{|V|}]$ ，关系表示记为 $\mathbf{R} = \{\mathbf{r}_{ij} \mid (i, j) \in E\}$ 。在图编码过程中，节点表示的更新不仅使用邻居节点的信息，也使用连接边上的关系信息。具体来说，在第 l 层中，模型先计算不同邻居节点对当前节点的重要程度，再结合邻居表示和边关系进行加权聚合，得到新的节点表示：

$$\alpha_{ij}^{(l)} = \text{softmax}_{j \in \mathcal{N}(i)} \psi(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}, \mathbf{r}_{ij}) \quad (3-10)$$

$$\mathbf{h}_i^{(l+1)} = \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} \phi(\mathbf{h}_j^{(l)}, \mathbf{r}_{ij}) \quad (3-11)$$

其中， $\psi(\cdot)$ 表示注意力打分函数，用于综合当前节点、邻居节点及边关系三者的信息； $\phi(\cdot)$ 表示关系感知的消息映射函数，用于生成沿边 (i, j) 传递的结构消息。经过 L 层传播后，得到节点的结构感知表示：

$$\mathbf{h}_i = \text{GraphEncoder}(\mathbf{X}, \mathbf{R}, G)_i \quad (3-12)$$

上述设计使图编码器同时建模节点之间的连接关系与边上的语义关系。经过图编码后，每个节点有三类表示可用于训练：图编码器输出的结构表示、节点自身的文本表示，以及由邻居节点文本聚合得到的上下文表示。具体地，节点 e_i

经图编码器得到节点图表示 $\mathbf{h}_i$ ，经文本编码器得到节点文本表示 $\mathbf{t}_i$ ；进一步对其邻居节点的文本表示进行聚合，可得到节点 e_i 的领域文本表示 $\mathbf{n}_i$ 。为了让节点表示同时保留自身语义和局部上下文信息，本文进一步构造三类对齐目标：节点图表示与节点文本表示的对齐、节点图表示与领域文本表示的对齐，以及节点文本表示与领域文本表示的一致性对齐。相应总损失可写为：

$$\mathcal{L}_{\text{stage1}} = \lambda_1 \mathcal{L}_{\text{g-t}} + \lambda_2 \mathcal{L}_{\text{g-d}} + \lambda_3 \mathcal{L}_{\text{t-d}} \quad (3-13)$$

其中各项均采用对比学习形式。首先，节点图表示与节点自身文本的对齐损失写为：

$$\mathcal{L}_{\text{g-t}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{t}_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{h}_i, \mathbf{t}_j)/\tau)} \quad (3-14)$$

其次，节点图表示与领域文本表示之间的对齐损失写为：

$$\mathcal{L}_{\text{g-d}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{n}_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{h}_i, \mathbf{n}_j)/\tau)} \quad (3-15)$$

最后，节点文本表示与领域文本表示之间的一致性对齐损失写为：

$$\mathcal{L}_{\text{t-d}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{t}_i, \mathbf{n}_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{t}_i, \mathbf{n}_j)/\tau)} \quad (3-16)$$

其中 $\text{sim}(\cdot, \cdot)$ 表示余弦相似度， τ 为温度系数。三项损失分别约束节点图表示对齐节点文本表示、对齐由邻居文本聚合得到的领域文本表示，并保持两类文本表示之间的一致性。经过这一阶段，模型得到能够同时表示事件语义、关系语义和局部图结构的节点表示，为下一阶段的因果归纳边预测提供输入。

第二阶段：因果归纳边预测。这一阶段的输入是事理图谱表示和待判断的事件对，输出是二者之间的因果归纳边强度。在获得第一阶段训练好的图编码器后，第二阶段将其输出作为结构 token 注入大语言模型，并在事件对文本查询的条件下预测因果归纳边的强度标签。具体地，对节点结构表示 $\mathbf{h}_i$ 进行线性投影，得到与语言模型隐空间对齐的图 token：

$$\mathbf{z}_i = W_p \mathbf{h}_i + \mathbf{b}_p. \quad (3-17)$$

随后，将事件图对应的全部节点 token $\{\mathbf{z}_i\}_{i=1}^{|V|}$ 以占位符替换的方式注入提示模板，并与事件对 (e_a, e_b) 的文本查询共同输入语言模型。第二阶段结合整个事件图的结构上下文，对事件对之间的因果强度进行预测。记语言模型最终用于分类的隐状态为 $\mathbf{o}$ ，则因果强度的条件分布可表示为：

$$p(y | G, e_a, e_b) = \text{softmax}(W_s \mathbf{o} + \mathbf{b}_s) \quad (3-18)$$

训练阶段采用 0–100 的细粒度因果强度标注作为监督信号，并在具体实现中将其映射为离散标签形式进行优化：

$$\mathcal{L}_{\text{stage2}} = - \sum_{c=0}^{100} \mathbb{I}(y=c) \log p(c \mid G, e_a, e_b) \quad (3-19)$$

上述实现既保留了因果强度的等级信息，也降低了小样本条件下直接进行连续数值回归可能带来的训练不稳定性。

综上，本节将事件对之间的因果归纳边强度预测定义为事理图谱上的因果归纳边补全任务。模型训练分为两个阶段：第一阶段学习事理图谱中事件和关系的表示，第二阶段利用这些表示学习给定事件对之间统计意义上的因果关联强度。该设计参考了 SAT 的基本思路，并针对本文的样本级事件图和因果强度预测任务进行了适配。

3.4.3 因果摘要生成与使用

这一阶段进一步解决的问题是，如何将上一节得到的统计因果关联强度预测组织为面向最终判断的整体依据。对于长文本情境中的因果判断，不同候选原因之间的相对影响、支撑路径以及替代解释关系仍需进行比较。为此，本文设计因果归纳模块，以校验后的事件图和因果归纳边补全结果为基础，对候选原因集合进行评分、路径恢复与排序，并在此基础上生成面向最终判断的结构化因果摘要，最后通过提示词方式用于辅助模型进行因果判断。

具体地，设校验后的事件图为 $\tilde{G} = (V, E)$ ，其中结果事件记为 e_b ，候选原因集合记为：

$$\mathcal{C} = \{e_a\} \cup \mathcal{E}_{\text{conf}} \quad (3-20)$$

其中 e_a 表示候选原因事件， $\mathcal{E}_{\text{conf}}$ 表示混淆事件集合。对于任一候选事件 $c \in \mathcal{C}$ ，将校验后的事件图 $\tilde{G}$ 、候选事件 c 以及结果事件 e_b 共同输入因果归纳边补全模型，得到候选事件 c 指向结果事件 e_b 的因果强度分数：

$$\hat{y}_{c \rightarrow b} = f(\tilde{G}, c, e_b), \quad \hat{y}_{c \rightarrow b} \in \{0, 1, \dots, 100\} \quad (3-21)$$

该分数对应事件图上节点对 (c, e_b) 的因果归纳边强度，用于衡量候选事件 c 在当前图结构中对结果事件 e_b 的因果贡献程度。在得到因果强度分数后，进一步在图 $\tilde{G}$ 上搜索从每个候选事件 $c \in \mathcal{C}$ 到结果事件 e_b 的可达事件链，并记所得路径集合为：

$$\mathcal{P}_{c \rightarrow b} = \{\pi \mid \pi : c \rightsquigarrow e_b\} \quad (3-22)$$

在得到每个候选事件的因果强度分数 $\hat{y}_{c \rightarrow b}$ 及其事件链集合 $\mathcal{P}_{c \rightarrow b}$ 后，按照分数从高到低对候选原因进行排序，并将排序结果记为 $\mathcal{R}$ 。排序结果 $\mathcal{R}$ 同时保留三类信息：候选事件本身、候选事件到结果事件的因果强度分数，以及支撑该

判断的事件链集合。由此，可在统一结构中比较候选原因事件与混淆事件集合中各事件对结果事件的相对影响，并提取与排序结果相对应的路径证据。

最后，将排序后的候选集合 $\mathcal{R}$ 作为结构化证据输入大语言模型，由其生成面向问答任务的因果摘要。该摘要综合比较候选原因事件与混淆事件在因果强度和事件链上的差异，并概括支持性证据与替代性解释。在此基础上，待测模型再依据因果摘要输出最终判断。由此，整条归纳流程可概括为：

$$\tilde{G} \xrightarrow{\text{score}(c, e_b)} \{\hat{y}_{c \rightarrow b}\}_{c \in C} \xrightarrow{\text{path search}} \{\mathcal{P}_{c \rightarrow b}\}_{c \in C} \xrightarrow{\text{ranking}} \mathcal{R} \xrightarrow{\text{LLM}} \text{因果摘要} \quad (3-23)$$

通过上述过程，图模型得到的统计因果关联强度预测被进一步组织为面向问题的整体因果归纳结果。因果归纳边、事件链与候选排序共同构成结构化比较证据，从而为后续语言模型的总结与决策提供更稳定、也更具可解释性的辅助信息。

最后，将该因果摘要与原始因果判断问题共同组成模型输入，使待测模型在因果摘要提供的候选排序、因果强度和事件链证据基础上输出最终判断。

3.5 实验

3.5.1 数据集

本节介绍实验所使用的数据集及其与本文任务设定的对应关系。训练阶段采用 CRAB 数据集构造因果归纳边补全模型的监督样本，测试阶段采用 MoCa 数据集评估所提出方法在长文本情境因果判断任务中的实际效果^[90,96]。为保持训练阶段与测试阶段在表示层面的统一性，本文在两类数据上采用一致的事理图谱抽取原则与关系类型体系。

CRAB 数据集提供长上下文背景、上下文中隐含的事件对，以及对应的 0–100 细粒度因果强度标注。该数据集能够为事件对之间的因果强度预测提供直接监督，因此作为本文因果归纳边补全模型的训练数据来源。围绕本文的任务设定，本文将 CRAB 中的原始样本重构为样本级事件图训练样本，具体过程如下。

(1) 首先，将原始上下文文本作为输入，围绕已标注事件对 (e_a, e_b) 进行问题驱动的事理图谱抽取，使图中显式包含候选原因事件与结果事件，并保留与二者相关的其他事件及其关系。

(2) 其次，对同一样本进行多次独立抽取，得到多个候选三元组集合，并按照前文定义的图级语义一致性得分选择代表图，以减小生成式抽取中的表述波动和局部结构缺失。

(3) 最后，将校验后得到的代表性事件图记为 G ，并将数据集给定的事件对与因果强度标注映射为 (e_a, e_b) 与监督标签 y ，从而构造得到 (G, e_a, e_b, y) 形式的训练样本。

经过上述处理，CRAB 数据集被转化为适用于本文任务的样本级事件图训练集。基于该训练集，模型能够在事件语义、关系语义与局部结构信息的联合约束下学习事件对之间的因果归纳边强度。

MoCa 数据集用于本文方法的测试。该数据集以较完整的长文本情境为输入,要求模型结合上下文对候选原因与结果之间的因果关系作出判断,能够较好反映长文本情境下因果归纳任务的实际难度。

表 3-1 展示了训练样本构建中使用的部分代表性事件间关系类型。

表 3-1 事理图谱关系类型
Table 3-1 Relation types in the eventic graph

类别	代表关系	含义
时间关系	先后、同时、持续、起始、结束	描述事件发生的先后顺序、同时性及时间边界
条件与约束关系	前提、许可、依赖、阻断	描述事件发生所依赖的条件、许可约束与阻断因素
过程与衔接关系	启动、推进、承接、阶段转换	描述事件在叙事过程中的展开方式及不同阶段之间的衔接关系
状态与影响关系	状态变化、增强、减弱、恢复、恶化	描述一个事件与另一事件在状态层面的影响和变化关系
组成与层级关系	组成、包含、子事件、并列环节	描述复合事件的内部结构及不同事件片段之间的层级关系
背景与情境关系	场景设定、环境约束、情境伴随	描述背景性事件、情境性事件与核心事件之间的关联方式
关联与解释关系	一般关联、迹象提示、证据支持、解释说明	描述较弱的语义连接,以及对事件判断有辅助作用的指示、证据与解释线索
共指与叙事推进关系	重述、共指、铺垫、转折、升级、对照	描述事件重述、叙事推进、情节转折及替代性情节对照关系

表中的关系类型能够为后续因果归纳提供关键约束信息。例如,当一个事件先于另一事件发生时,这一时序关系可以为因果方向提供基本约束;当一个事件构成另一事件发生的前提,或对其形成阻断时,这类条件与约束关系有助于区分支持性路径与排除性路径;当一个事件引出新的状态变化,或对后续情节提供解释线索时,这类过程与解释关系能够补充事件演化链条,为候选原因排序提供结构证据。由此,这些非直接因果关系共同构成了面向因果归纳的图结构基础。

3.5.2 模型构建和训练

本节说明实验中的模型具体构建方式与训练流程。整体上,模型训练分为两个阶段:第一阶段学习结构感知节点表示,第二阶段在全图结构上下文与事件对查询条件下预测因果强度。

第一阶段用于训练事件图编码器。实验中采用 Sentence-Transformers^[97] 的 all-mpnet-base-v2 对事件节点文本和事件间关系文本进行初始化编码,并以 Graph-Transformer 作为图编码器。图编码器设置为 3 层、8 个注意力头,隐藏维度为 512。训练目标对应前文定义的三类对齐损失,即节点图表示、节点文本表示和领域文本表示在统一语义空间中的联合对齐。为降低训练波动并保留预训练语

义空间,该阶段默认冻结文本编码器,仅更新图编码器及相关投影层。训练时共进行 50 个 epoch,学习率设为 2×10^{-5} ,batch size 设为 8,邻居采样数设为 3。第一阶段训练完成后,保存图编码器参数,供第二阶段直接加载使用。

第二阶段用于预测事件对之间的因果归纳边强度。实验中采用 Qwen2.5-1.5B 作为下游语言模型,并加载第一阶段训练完成的图编码器参数。首先由图编码器生成事件图中各节点的结构表示,再通过投影模块将这些表示映射到语言模型隐空间,并组织为与图结构对应的结构 token;随后,结构 token 与事件对 (e_a, e_b) 的文本查询共同构成模型输入,使语言模型能够在统一输入序列中同时接收图结构信息与任务查询信息。该阶段最大输入长度设为 512,输出类别数设为 101,对应 0-100 的细粒度因果强度标签。训练时共进行 8 个 epoch,学习率设为 1×10^{-4} ,batch size 设为 4。在此基础上,模型在全图结构上下文中完成因果强度预测,并由分类头输出 0-100 的潜在因果关系估计。

3.5.3 实验结果与分析

实验中设置了多类模型输入形式。Baseline 表示直接将原始长文本情境输入模型并输出判断结果。思维链 (Chain of Thought, CoT) 表示在原始文本基础上加入分步推理提示。一致性思维链 (Chain of Thought with Self-Consistency, CoT-SC) 表示对同一问题生成多条推理链,并通过答案聚合得到最终判断^[63]。思维树 (Tree of Thoughts, ToT) 表示生成多个候选推理分支,并对分支进行搜索和选择^[3]。思维图 (Graph of Thoughts, GoT) 表示用图结构组织中间推理结果,使不同推理节点可以组合和复用^[66]。CausalCoT^[4] 表示在原始文本基础上加入面向因果任务的结构化提示。本文方法包括事件摘要 (Event Summary, ES) 和因果摘要 (Causal Summary, CS) 两种输入形式。ES 是本文方法的消融变体,表示仅将事理图谱抽取结果整理为事件结构摘要后输入模型;CS 是本文的完整方法,表示在事件结构基础上进一步结合事理图谱补全、因果边补全和候选路径比较生成因果摘要。

本节对实验结果展开分析。首先,分析本文方法在长文本因果判断任务中的整体效果,并结合消融结果比较事件结构摘要与因果摘要两类中间表示的作用差异。随后,考察不同模型规模下各方法的性能变化,以分析本文方法在不同模型能力水平上的收益表现。最后,从细粒度因果因素层面对不同方法进行比较,以更具具体地说明性能差异所对应的因果判断能力来源。

3.5.3.1 本文方法的整体效果与消融实验分析

下面,本文从整体效果和消融结果两个方面分析本文方法在长文本因果判断任务中的作用。

表 3-2 展示了 Qwen 系列模型在 Baseline、CoT、CoT-SC、ToT、GoT、CausalCoT 以及本文完整方法 CS 等不同方法设置下的因果判断的正确率。从结果来看,CoT 在三种模型上均带来一定提升,但提升幅度有限。CoT-SC 的收益较弱,在 1.5B 和 30B 模型上低于普通 CoT,说明多次采样和答案聚合难以稳定解决长文本因果

判断中的证据选择问题。ToT 是新增方法中表现最好的对比方法，在 4B 与 30B-A3B 模型上分别达到 61.1 和 66.7，说明搜索多个候选推理分支能够帮助模型比较不同解释。GoT 在部分模型上也有提升，但随模型规模变化较大。CausalCoT 在 1.5B 模型上优于 CoT，在 4B 与 30B-A3B 模型上低于 Baseline。CS 在三种模型规模上均取得最高结果，分别达到 67.0、73.0 与 75.6，相对最优对比方法分别提升 13.5、11.9 与 8.9 个百分点。

结合不同方法的结果可以看到，通用推理增强方法能够改善部分模型的表现，但收益不稳定。CoT-SC 依赖多条推理链之间的答案一致性，ToT 依赖候选分支搜索，GoT 依赖中间推理节点的组织和复用。这些方法主要调整模型生成推理过程的方式。对于长文本因果判断任务，模型仍需要自行从原始文本中寻找关键事件、排除干扰事件，并比较不同候选原因的作用。当模型已经受到叙事顺序、局部显著情节或常识联想影响时，多路径采样、分支搜索和图式组织也可能围绕相同的表层线索展开，因此收益会随模型规模和具体方法发生波动。

CS 的作用更集中在判断依据的整理上。本文方法先抽取事理图谱，将长文本中的事件、条件、状态变化、阻断关系和解释线索整理为图结构；随后补全候选原因到结果事件的因果边，得到候选原因的因果强度；最后结合路径搜索和候选排序生成因果摘要。经过这一过程，模型输入中的关键信息从原始长文本转为候选原因、结果事件、支持路径和因果强度。对于较小模型，因果摘要降低了在长文本中查找证据的难度；对于较大模型，因果摘要减少了叙事顺序和显著情节对判断的干扰。模型最终依据经过筛选的因果摘要进行判断，因此在三个模型规模上都取得更高准确率。实验结果表明，面向因果判断的结构化证据能够为长文本归因提供更稳定的支持。

表 3-2 不同方法设置下的因果判断准确率比较

Table 3-2 Causal Judgment Accuracy under Different Method Settings

方法	Qwen2.5-1.5B	Qwen3-4B	Qwen3-30B-A3B
Baseline	45.1	55.0	61.0
CoT	51.3	56.0	62.5
CoT-SC	45.8	56.3	61.1
ToT	49.3	61.1	66.7
GoT	50.7	55.6	63.2
CausalCoT	53.5	49.0	54.1
CS (ours)	67.0 (13.5↑)	73.0 (11.9↑)	75.6 (8.9↑)

注：CS 表示本文完整方法生成的 Causal Summary；括号内为相对对应模型下最优对比方法的提升。

消融实验。图 3-3 展示了 Baseline、ES 和 CS 三种输入形式在 Qwen 系列模

型上的比较。实验证明仅进行事理图谱摘要，能够在一定程度上减弱长文本中的冗余干扰，并在 1.5B 与 4B 模型上分别提升至 56.3 和 60.4，整体优于对应的 Baseline，说明事件结构整理本身已经能够为因果判断提供一定帮助。这表明，将原始长文本先归纳为事理图谱并进一步压缩为事件摘要，能够为模型提供比开放式提示更集中的结构信息。然而，ES 在 30B-A3B 模型上下降至 56.25，低于对应 Baseline，这说明一般事件结构摘要本身仍不足以稳定支撑最终因果判断；随着模型自身能力增强，模型已经能够在一定程度上从原始文本中掌握较好的非因果关系理解，事件结构摘要甚至会引入干扰项，从而对模型的因果判断造成负面影响。

在各模型规模下，CS 均表现出相对于 ES 的稳定优势。该结果说明，本文方法带来的性能提升并不主要依赖对原始文本的概括性压缩，而更多来自事理图谱补全后对因果相关结构的针对性筛选与重组。ES 所保留的是较一般的事件结构信息，能够在一定程度上减弱原始文本中的冗余干扰，但仍不足以稳定支持最终判断；CS 则在事件图的基础上进一步结合因果边补全结果，对候选原因与结果事件之间的支持路径进行显式比较，并将更具判别力的因果相关结构压缩为输入表示。

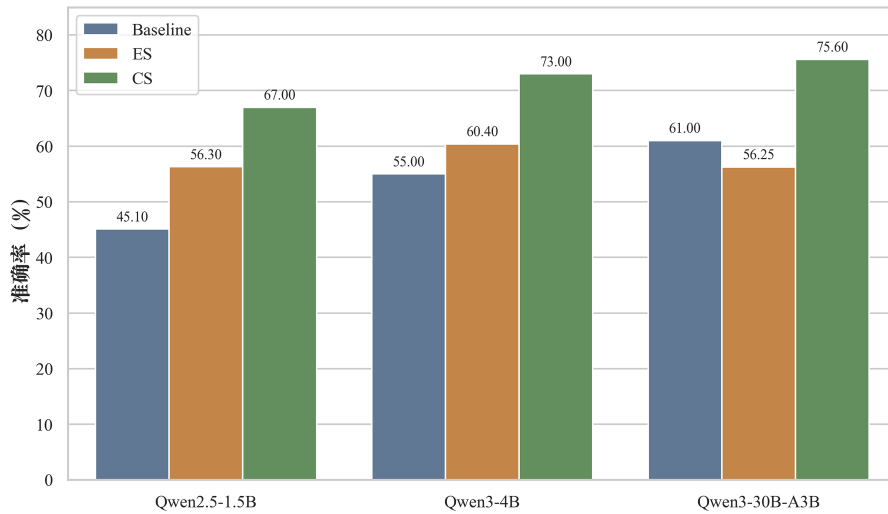


图 3-3 事件摘要与因果摘要的消融结果比较

Figure 3-3 Ablation Results for Event and Causal Summaries

实验结果表明，本文方法在不同模型规模下均具有稳定效果，其中，事件结构整理为因果判断提供了基础支撑，而事理图谱补全基础上的因果归纳过程进一步提升了模型对关键因果证据的利用能力。

3.5.3.2 模型规模的影响

图 3-4 展示了不同模型规模下不同方法设置对模型因果判断能力的影响。

在较小的模型上，不同方法的增益差异明显。以 Qwen2.5-1.5B 为例，CoT、CoT-SC、ToT、GoT、CausalCoT 和 CS 分别带来 6.2、0.7、4.2、5.6、8.4 和 21.9 个

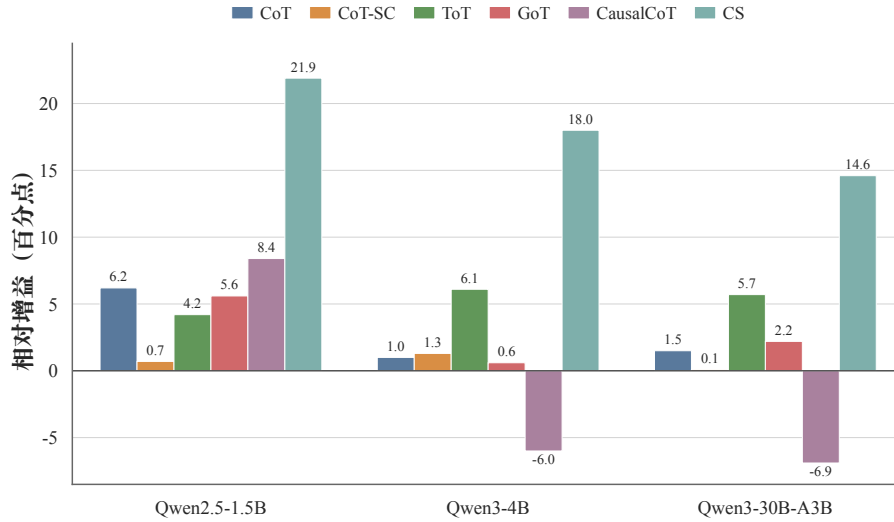


图 3-4 不同模型规模下各方法的相对增益

Figure 3-4 Relative Gains by Method across Model Scales

百分点的提升。CoT-SC 的提升较小，说明多次采样和答案聚合对该任务的帮助有限。CoT、ToT、GoT 和 CausalCoT 能够带来一定收益，说明显式推理提示、分支搜索和中间节点组织可以帮助小模型整理部分线索。CS 的提升幅度最大，说明因果摘要能够直接提供候选原因、结果事件和支持路径之间的关键信息。

随着模型规模增长到 4B 及以上，不同方法的收益继续分化。图 3-4 显示，CoT 在 4B 和 30B-A3B 上仅带来 1.0 和 1.5 个百分点的提升，CoT-SC 的提升分别为 1.3 和 0.1 个百分点。GoT 在 4B 上提升 0.6 个百分点，在 30B-A3B 上提升 2.2 个百分点，整体波动较大。ToT 在两个较大模型上表现较好，分别提升 6.1 和 5.7 个百分点，说明分支搜索对较强模型有帮助。CausalCoT 在两个较大模型上出现下降，分别低于 Baseline 6.0 和 6.9 个百分点，说明复杂的因果步骤提示可能引入错误中间过程。

CS 在三个模型规模上都保持最高增益，分别提升 21.9、18.0 和 14.6 个百分点。该结果说明，因果摘要在不同规模模型上都能提供稳定帮助。对于较小模型，因果摘要降低了从长文本中寻找关键线索的难度。对于较大模型，因果摘要把判断依据集中到候选原因、结果事件、因果强度和支持路径上，减少了模型沿表层线索作答的情况。

3.5.3.3 细粒度因果因素分析

前文已经比较了不同方法在整体表现和不同模型规模下的收益差异，为进一步探究这些差异具体来源于哪些因果判断能力，本节进一步从细粒度因果因素层面对不同方法进行分析，以考察不同方法在不同因果因素上的响应模式。本文所考察的因素包括：

- (1) **后发原因与先发原因**：反映时间顺序对因果判断的影响。
- (2) **异常事件与正常事件**：区分某个候选事件相对于故事中给定规则、常态

或功能预期是更异常还是更正常。

(3) **作为与不作为**：区分结果是由某个行动直接引起，还是由于某个本应发生的行动没有发生而引起。

(4) **规范性规则与统计性常态**：区分故事中的“常态”是来自明确规定的行为规则，还是来自通常情况下更常见的经验模式。

(5) **知情与不知情**：表示行动者是否知晓其行为的异常性或后果。

图 3-5 汇总了三个 Qwen 模型在上述因素上的表现，用以分析整体性能差异在具体因果因素上的分布。

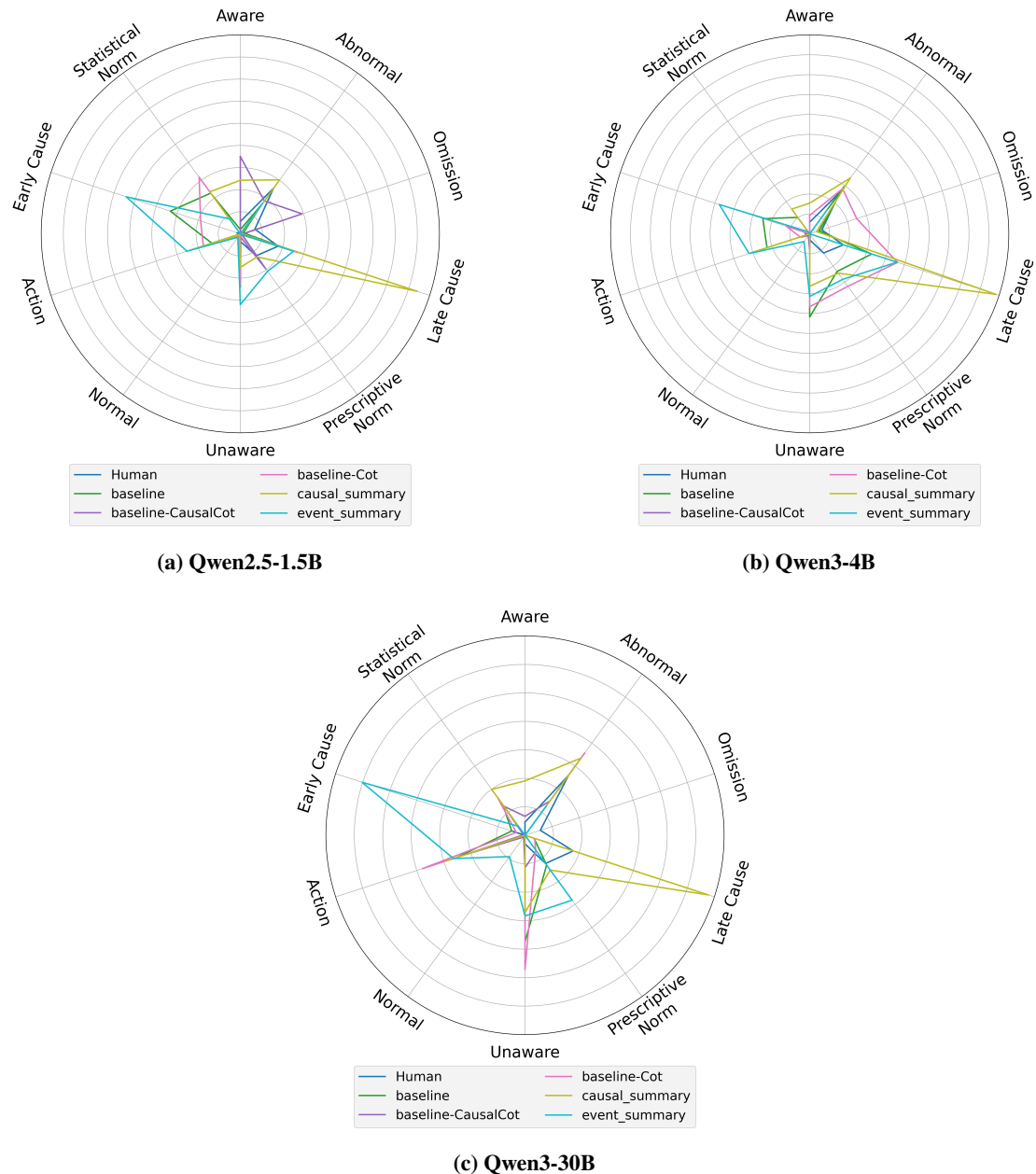


图 3-5 不同方法在细粒度因果因素上的响应

Figure 3-5 Sensitivity of Different Methods to Fine-Grained Causal Factors

从图 3-5 可以看出，不同方法的增益集中体现在若干结构性较强的因果维度

上。其中，后发原因与先发原因这一因素上的差异最为明显。在三个模型规模上，CS 在该维度上的响应均明显高于其余方法，说明该表示形式更有利于模型利用时间顺序信息进行因果判断。结合前文方法部分可以理解为，事件图中的时序关系及其路径结构为模型提供了比原始叙事更清晰的方向性约束，而这些约束在因果摘要中得到进一步保留与强化，从而增强了模型在后发原因与先发原因区分上的响应强度。

异常性相关因素构成了另一类较为稳定的改进来源。在一般“异常事件与正常事件”维度上，CS 在 1.5B 与 4B 模型上响应较强，在 30B 模型上则与 Baseline 和 CoT 接近。该结果表明，经过事件图归纳、因果边补全与因果摘要组织之后，模型在面对正常事件与异常事件并存的情形时，更容易把偏离故事常态的事件判断为结果发生的关键原因。

与此不同，ES 的响应模式更多体现为对一般事件关系的组织与压缩，而非对关键归因因素的集中增强。从图中可以看到，在“后发原因与先发原因”和“异常事件与正常事件”等维度上，ES 的响应整体弱于 CS。由此可见，ES 更有利于模型把握事件之间的一般关系，而在关键致因事件识别方面所提供的支持相对有限。

从图中的其余维度可以看到，行为者意识、作为与不作为以及规范类型等因素在不同模型和不同方法之间普遍波动较大，尚未呈现出稳定一致的响应模式。在“知情与不知情”维度上，多数方法与人工趋势之间仍存在偏离；在“作为与不作为”维度上，不同方法的响应方向也不完全一致；在“规范性规则与统计性常态”维度上，各方法之间未形成稳定的分布层次。这些结果表明，行动者知情状态、作为与不作为差异以及规范解释等因素，目前尚难通过图结构中的事件关系与路径信息得到充分刻画。

综合图 3-5 可以看出，不同方法在各细粒度因果因素上的响应存在明显差异。其中，后发原因与先发原因以及异常性相关因素上的变化最为明显；ES 的作用则主要体现为对一般事件关系的组织；知情与不知情、作为与不作为以及规范性因素仍然是当前方法波动较大的维度。由此可见，本文方法在结构性较强、可由事件关系和图路径显式表达的因果因素上取得了较为明显的改进，而在更高层次的知情状态和规范解释因素上仍有进一步提升空间。

3.6 小结

本章围绕长文本情境下的因果判断任务，提出了一种基于事理图谱补全的因果推理方法。针对大语言模型在原始文本中容易受到冗余描述、局部显著信息和表层相关模式干扰的问题，本文将因果判断过程组织为“事理图谱抽取与校验-因果归纳边补全-因果摘要生成”三个阶段。该方法首先通过问题驱动的事理图谱抽取与一致性校验，将长文本中的事件及其依赖关系转化为较稳定的结构化表示；随后在事件图上建模候选事件对结果事件的潜在因果贡献，并结合路径搜索与排序形成结构化因果证据；最后生成面向最终判断的因果摘要，为下游模

型提供更集中的判断依据。

本文进一步构建了面向样本级事件图的因果归纳边补全模型。该模型以事件节点文本和关系文本的联合表示为基础，通过结构感知图编码器学习节点表示，并在全图上下文中预测候选事件对的因果强度。通过将时序、条件、阻断、状态变化和组成关系等非因果结构信息纳入因果归纳过程，模型能够在结构空间中组织和比较候选证据，从而增强对关键因果路径的识别能力，为后续候选原因排序与因果摘要生成提供支撑。

实验结果表明，现有通用模型在长文本因果判断任务上的基础表现仍然有限。CoT-SC、ToT 和 GoT 等通用推理增强方法在部分模型上有效，但收益随方法和模型规模变化。进一步分析表明，本文方法生成的因果摘要在不同模型规模下均表现出稳定有效性，说明将原始文本中的相关性判断转化为围绕候选原因与结果事件的因果结构比较，能够为最终判断提供更有效的支持。消融结果进一步表明，事件结构整理能够为因果判断提供基础支撑，而事理图谱补全基础上的因果归纳过程进一步增强了模型对关键因果证据的利用能力。模型规模分析显示，随着模型规模增大，不同方法的收益呈现明显分化，因果摘要始终保持最高正增益；细粒度因果因素分析则表明，本文方法的改进主要体现在后发原因与先发原因、异常事件与正常事件等结构性较强的因果因素上。整体而言，本章工作验证了通过事理图谱补全提升大语言模型长文本因果判断效果的有效性。

第4章 基于过程验证的反事实因果推理方法

4.1 引言

第三章围绕长文本情境中的因果判断问题展开，通过事理图谱补全与因果归纳边建模，推动模型由对表层相关模式的利用转向对因果关系的结构化判断。在因果判断基础上，进一步分析条件变化后的结果演化，是更高层次的反事实因果推理问题。因果判断主要回答的是某一因素是否构成结果发生的原因，尚不能进一步说明当条件发生变化时结果将如何演化。进一步分析“若条件改变，结果将如何变化”，需要模型具有较好的反事实因果推理能力。因此，本章在前一章工作的基础上，进一步研究反事实因果推理方法。

这一研究也来源于广泛存在的决策分析需求。在金融风控、医疗决策、公共治理和经营管理等场景中，在完成对结果成因的判断之后，研究者和决策者通常还需要评估当某一条件发生调整时结果将如何变化。例如，在医疗决策中，当已有分析表明某一治疗方案与患者当前结果存在关联之后，进一步的问题往往是：若改变治疗方案、剂量或干预时点，患者状态将如何演化。这类围绕条件变化展开的结果分析，构成了反事实推理在现实决策中的典型问题形式。

近年来，围绕大语言模型反事实能力的研究持续推进。相关工作表明，大语言模型在反事实因果推理任务中，表现仍然不够稳定^[85,86,88]。其中一个突出问题在于，模型常常跳过必要推演过程，直接依据局部线索生成结果，因而难以分析条件变化后的结果。已有研究在可执行反事实任务的推理轨迹分析中也观察到类似现象：模型面对数学反事实问题时，对于影响计算结果的未知变量，模型常出现枚举未知变量、任意假设未知变量以及不必要分情况讨论等失败模式。^[6]

反事实推理的一般形式是：在给定某个原因的情况下，得出了一个确定的结果，同时存在一些未知量会影响原因和结果的对应关系，要求分析当原因改变时，结果怎么变化。反事实推理具有清晰的推理过程：首先是根据给定已知的信息分析出未知量。其次，在保持未知量不变的情况下，知道原因怎么改变。最后，根据分析出的未知量和新的原因推导出结果的变化。

基于上述背景，本章提出一种基于过程验证的反事实因果推理方法。第一是构造符合反事实推理过程的训练数据对模型进行监督微调，使模型学习更清晰的反事实求解过程。第二是针对反事实推理过程设计奖励信号，通过强化学习增强模型对反事实推理过程的学习。

本章其余内容安排如下：首先对本章关注的研究问题进行说明；随后介绍与本章相关的研究工作；在此基础上依次说明基于过程验证的反事实因果推理方法；最后通过实验验证本章方法的有效性。

4.2 研究问题

本章关注反事实推理任务。给定一个事实情境及其已观测结果，并在此基础上引入某一条件变化，模型需要判断在保持事实基础一致的前提下，替代情境中的结果将如何变化。这类问题往往存在会影响判断的未知量，本文统一称为隐藏状态。

下面给出一个说明性示例。

任务示例：已知在当前事实情境下，珠穆朗玛峰是世界最高峰。若地壳运动使珠穆朗玛峰的高度降低 300 米，而其他相关地理条件保持不变，那么世界最高峰将变为哪一座山？

对于这一问题，反事实推理可以分为三个步骤。第一步是根据题目给出的已有事实推断隐藏条件。例如，由“珠穆朗玛峰是世界最高峰”这一事实，可以进一步去获得珠穆朗玛峰当前高度约为 8848.86 米，世界第二高峰为乔戈里峰，其高度约为 8611 米等未显示给出的条件。第二步是对题目指定条件施加变化，即将珠穆朗玛峰高度降低 300 米并保持其他条件不变。第三步是比较变化后的珠穆朗玛峰高度与乔戈里峰高度，并据此判断反事实情境下的世界最高峰。

为对反事实问题进行统一描述，本文采用带隐藏状态的形式化表示。设 x 表示可观测输入， U 表示不可直接观测的隐藏状态， y 表示在给定输入与隐藏状态下得到的输出，则事实过程可表示为：

$$y = f(x, U) \quad (4-1)$$

其中 $f(\cdot)$ 表示由输入与隐藏状态共同决定输出的确定性映射关系。对于同一问题，事实观测通常给出输入 x 与对应输出 y ，但并不直接揭示隐藏状态 U 的具体取值。

在此基础上，反事实问题可表述为：已知事实输入 x 下观测到输出 y ，若将输入替换为新的反事实输入 x' ，并保持事实发生时的隐藏状态不变，则可能得到怎样的输出。由此，与事实观测一致的隐藏状态集合可定义为：

$$\mathcal{U}(x, y) = \{U \mid f(x, U) = y\} \quad (4-2)$$

该集合刻画了所有能够解释当前事实观测的潜在状态。进一步地，在反事实输入 x' 下，对应的反事实输出集合可写为：

$$\mathcal{Y}_{cf}(x', x, y) = \{f(x', U) \mid U \in \mathcal{U}(x, y)\} \quad (4-3)$$

若 $\mathcal{U}(x, y)$ 只包含唯一隐藏状态，则反事实输出唯一；若 $\mathcal{U}(x, y)$ 含有多个与事实一致的潜在状态，则反事实输出也可能表现为一个结果集合。

反事实推理的难点主要体现在三个方面。第一，模型需要先根据题目给出的事实结果，推回与之相符的隐藏状态，而这些信息通常并没有被直接写出。第

二，在构造反事实情境时，模型只能修改题目明确指定的条件，其他事实基础必须保持不变，需要模型理解并执行关键条件的更新。第三，在完成条件替换之后，模型还需要基于保持不变的事实基础，继续推导结果会如何变化。

现有大语言模型在处理这类任务时，往往不能清楚地区分这几个步骤。它们常常把隐藏状态推断、条件替换和结果推导混在同一段连续生成中，甚至直接跳过中间过程，根据局部线索猜测答案。这样一来，模型就容易出现三类问题：一是没有正确推断事实约束，二是错误改动了本应保持不变的条件，三是在替代情境中的结果推导不完整或不一致，得出错误的分析结果。

针对上述问题，本文提出基于过程验证的反事实因果推理方法，以提升反事实推演过程的稳定性。

4.3 相关工作

围绕大语言模型反事实推理的研究，现有工作首先从能力评测与任务构造出发，逐步揭示模型在这一方向上的主要挑战。早期研究多从相关性判断、因果方向识别或常识性因果问答切入，用以考察模型是否能够区分变量之间的因果关系；随后，评测对象进一步扩展到干预推理和反事实推理，开始关注模型能否在给定事实观测的基础上处理替代情境下的结果变化^[4,85,86,88]。相关结果表明，大语言模型在因果反事实任务上仍存在依赖猜测、跳过必要步骤等缺陷。

除能力评测与任务构造外，现有研究也从训练层面探索反事实因果推理能力的提升路径。具体而言，一类方法围绕事实问题与反事实问题之间的对应关系设计训练反馈，通过构造事实-反事实配对样本，并结合因果一致性导向的优化目标，提升模型在因果相关问题中的表现与一致性^[76]。另一类方法则将因果规则以显式形式编码进训练数据，通过构造〈前提、假设、结论〉式样本，将传递性、d-separation等规则直接纳入模型训练过程，使模型能够从结构化示例中学习因果约束^[77]。这些研究表明，针对反事实场景构造训练信号，有助于对模型的因果与反事实推理能力形成更明确的引导。

除此之外，已有研究还进一步将反事实问题组织为代码或形式化数学任务。^[6]这类反事实任务具有清晰的推理过程，为反事实研究提供了更稳定的任务基础，也使研究者能够在更受控的条件下考察模型处理反事实问题的行为特征。总体来看，现有研究已经从评测、训练与任务设计等多个方面推动了大语言模型反事实推理研究的发展。

4.4 方法介绍

针对反事实因果推理过程中模型推理过程边界不清晰、跳过关键步骤或关键步骤推理失败的问题，本文提出基于过程验证的反事实推理方法。

在反事实推理中，归因是指推断与事实结果相符的隐藏条件，干预是指在其他条件不变时替换指定条件，预测是指推到出反事实情境下的结果。

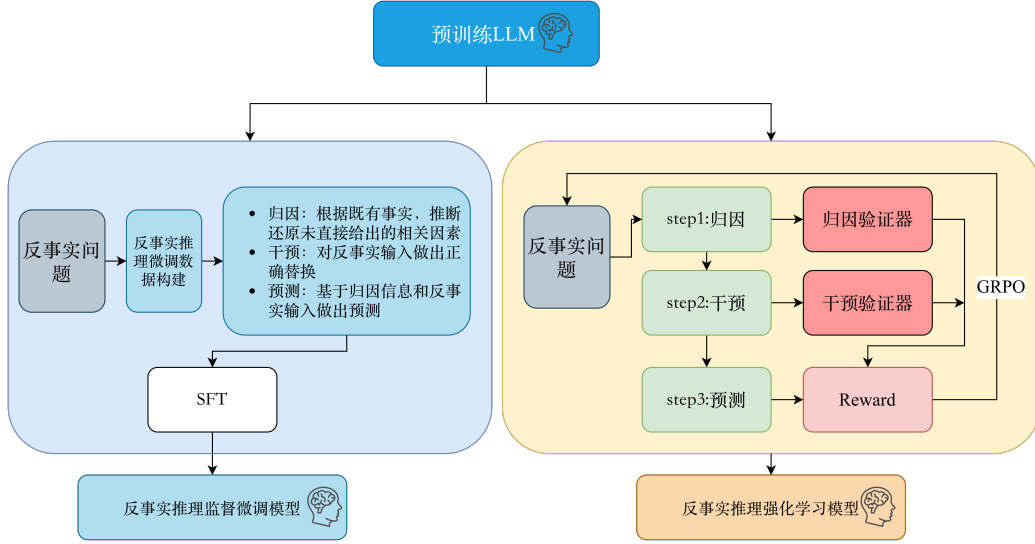


图 4-1 基于过程验证的反事实因果推理方法

Figure 4-1 Counterfactual causal reasoning method based on process verification

本文的方法主要包括两部分。如图 4-1 所示，第一部分基于三阶段推理数据的监督微调方法，通过构造完整覆盖反事实推理过程的数据，使模型学习更清晰的反事实求解过程，得到反事实推理监督微调模型。第二部分是强化学习，通过设计归因验证器和干预验证器两个奖励模型，为模型提供过程级验证信号，并据此增强关键中间过程的学习，得到反事实推理强化学习模型。

4.4.1 基于三阶段推理数据的监督微调方法

本文通过构造面向反事实因果推理过程的归因-干预-预测三阶段推理数据，并通过监督微调使模型从训练样本中学习更清晰的反事实推理过程。经过该方法训练后得到的模型，本文称为反事实推理监督微调模型。记事实输入为 x ，事实输出为 y ，反事实输入为 x' ，隐藏状态为 U 。事实世界与反事实世界可分别写为：

$$Y_x(U) = f(x, U), \quad Y_{x'}(U) = f(x', U) \quad (4-4)$$

由此，与事实观测 (x, y) 一致的上下文集合和对应的反事实输出集合分别为：

$$\mathcal{U}(x, y) = \{U \mid Y_x(U) = y\}, \quad \mathcal{Y}_{cf}(x', x, y) = \{Y_{x'}(U) \mid U \in \mathcal{U}(x, y)\} \quad (4-5)$$

记归因、干预、预测三个阶段对应的目标输出分别为 a 、 i 和 p ，则完整目标可记为：

$$S = (a, i, p) \quad (4-6)$$

其中， a 对应事实一致上下文 $\mathcal{U}(x, y)$ 的推断， i 对应共享事实基础上对输入条件实施 $x \rightarrow x'$ 的替换， p 对应在该替换条件下得到的反事实结果 $\mathcal{Y}_{cf}(x', x, y)$ 的推导与最终预测。监督微调的作用在于，使模型在结构化监督下学习归因、干

预和预测三个阶段所对应的反事实推理行为。相应地，监督微调损失写为：

$$\mathcal{L}_{\text{SFT}} = -\log p_{\theta}(S | q) \quad (4-7)$$

该目标同时对最终答案和具有明确过程结构的输出序列进行建模。由此，监督微调能够通过三阶段推理样本引导模型形成更清晰的反事实推理过程。

4.4.2 基于验证器信号的强化学习方法

除监督微调之外，本文还采用 Group Relative Policy Optimization (GRPO) 对模型进行强化学习优化。该设置同时使用答案正确性奖励、归因验证器提供的过程信号以及干预验证器提供的过程信号，引导模型形成更稳定的反事实推理行为。经过该方法训练后得到的模型，本文称为反事实推理强行学习模型。归因验证器与干预验证器的具体结构将在后文过程级验证模块中进一步说明。

对于给定输入 q ，从当前策略 π_{θ} 中采样一组候选输出：

$$\{\hat{S}_i\}_{i=1}^G, \quad \hat{S}_i \sim \pi_{\theta}(\cdot | q) \quad (4-8)$$

其中每个 $\hat{S}_i$ 都是模型针对输入 q 生成的候选输出。对每个候选输出，构造奖励函数：

$$R(q, \hat{S}_i) = \alpha_1 R_1(q, \hat{S}_i) + \alpha_2 R_2(q, \hat{S}_i) + \alpha_3 R_3(q, \hat{S}_i) \quad (4-9)$$

其中， R_1 表示答案正确性奖励， R_2 表示归因行为奖励， R_3 表示干预行为奖励。后两项奖励分别由归因验证器与干预验证器提供，用于针对归因过程与干预过程进行专门判别，并为模型提供更有针对性的过程监督信号。

在得到一组候选输出的奖励之后，GRPO 采用组内标准化方式构造相对优势。设第 i 个候选的奖励为 $r_i = R(q, \hat{S}_i)$ ，则其相对优势定义为：

$$A_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r}) + \epsilon}, \quad \mathbf{r} = (r_1, r_2, \dots, r_G) \quad (4-10)$$

对于第 i 个候选输出，策略比率定义为：

$$\rho_i = \frac{\pi_{\theta}(\hat{S}_i | q)}{\pi_{\theta_{\text{old}}}(\hat{S}_i | q)} \quad (4-11)$$

于是，GRPO 的优化目标可写为：

$$\begin{aligned} \mathcal{L}_{\text{GRPO}} = & -\frac{1}{G} \sum_{i=1}^G \min [\rho_i A_i, \text{clip}(\rho_i, 1 - \epsilon_c, 1 + \epsilon_c) A_i] \\ & + \gamma \text{KL}(\pi_{\theta}(\cdot | q) \| \pi_{\text{ref}}(\cdot | q)) \end{aligned} \quad (4-12)$$

其中 ϵ_c 表示截断阈值， π_{ref} 表示参考策略， γ 表示 KL 正则项的权重。总体而言，该强化学习设置通过结合验证器信号与答案奖励，对模型的反事实推理行为进行优化，从而进一步增强模型在归因、干预等关键过程上的表现。

4.4.3 过程级验证模块

本文进一步设计过程级验证模块，用于判断强化学习中模型生成的中间过程是否满足反事实因果推理步骤。该模块包括归因验证器和干预验证器，这两个奖励模型用于监督模型的强化学习过程，引导模型更稳定地学习反事实因果推理模式。

4.4.3.1 归因验证器

归因验证器用于判断模型在归因阶段中是否完成了由根据已有事实推断隐藏状态。归因验证器的结构如4-2所示。

给定题目文本和已有事实和模型的归因推理过程。将模型的归因推理过程记为候选归因文本。设题目文本为 q ，事实观测为 o ，候选归因文本为 a ，则验证器输入记为：

$$z_a = [q; o; a] \quad (4-13)$$

在该输入上，验证器输出一个整体分数和三个槽位分数，分别记为：

$$\mathbf{p}_a, \quad p_o, \quad p_h, \quad p_r \quad (4-14)$$

其中， $\mathbf{p}_a$ 评估候选归因文本是否包含良好的归因推理过程； p_o 表示文本中是否明确引用了事实； p_h 表示文本中明确讨论了隐藏状态； p_r 表示文本中是否进行隐藏状态的推断、筛选或排除。通过这一输出形式，验证器能够细粒度的评估模型反事实推理过程中的归因行为。

在模型结构上，归因验证器采用共享编码器与槽位级注意力聚合相结合的方式。设编码器输出的候选归因片段表示为：

$$H_a = [h_1, h_2, \dots, h_n] \quad (4-15)$$

其中 n 为候选归因文本的 token 数。为了分别刻画事实观测、隐藏状态提及与隐藏状态推断三个槽位，本文引入三个查询向量，分别记为 u_o 、 u_h 和 u_r 。以任一槽位 $k \in \{o, h, r\}$ 为例，其注意力权重定义为：

$$\alpha_{k,t} = \frac{\exp(u_k^\top h_t)}{\sum_{j=1}^n \exp(u_k^\top h_j)} \quad (4-16)$$

据此可得该槽位的聚合表示：

$$v_k = \sum_{t=1}^n \alpha_{k,t} h_t \quad (4-17)$$

三个槽位表示分别经过对应的分类头，得到槽位分数：

$$p_k = \sigma(W_k v_k + b_k), \quad k \in \{o, h, r\} \quad (4-18)$$

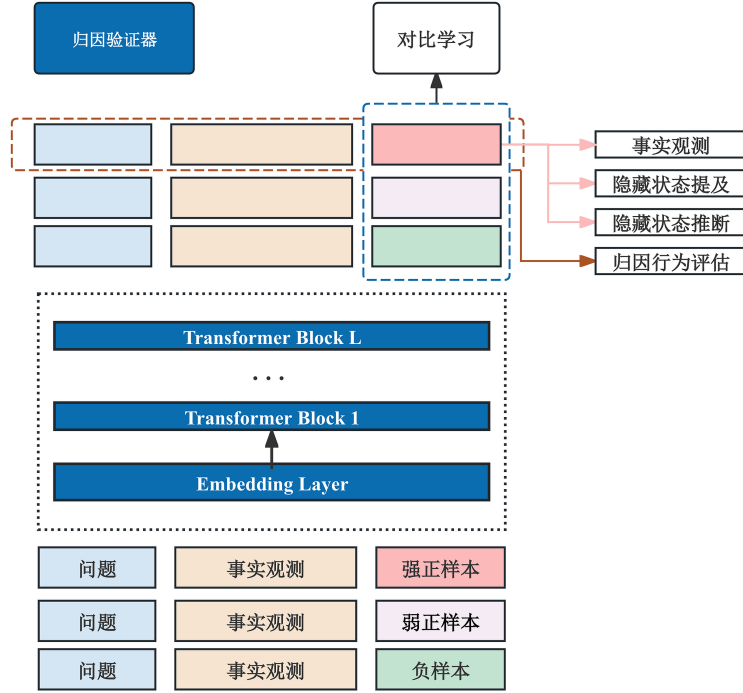


图 4-2 归因验证器结构示意图

Figure 4-2 Architecture of the abduction verifier

其中 $\sigma(\cdot)$ 表示 Sigmoid 函数。进一步地，记编码器的全局表示为 c_a ，则整体归因分数写为：

$$\mathbf{p}_a = \text{softmax} (W_a[c_a; v_o; v_h; v_r] + b_a) \quad (4-19)$$

归因验证器的训练依赖分组构造的监督样本。对于同一道题目，本文围绕同一题目文本 q 与事实观测 o 构造多条候选归因文本，根据归因过程的正确性赋予不同层次的标签。

在训练目标上，本文同时使用分类监督与对比学习监督。分类部分包括整体标签损失与槽位标签损失。设整体标签为 y_a ，三个槽位标签分别为 y_o 、 y_h 和 y_r ，则分类损失写为：

$$\mathcal{L}_a^{\text{cls}} = \text{CE}(\mathbf{p}_a, y_a) + \text{BCE}(p_o, y_o) + \text{BCE}(p_h, y_h) + \text{BCE}(p_r, y_r) \quad (4-20)$$

其中，整体标签采用类别标注，用于刻画候选归因文本在整体层面的判别结果；槽位标签采用二值标注，用于刻画各局部成分是否成立。为了进一步拉开不同质量归因文本之间的表示距离，本文在组内引入对比学习。设同组内强正样本表示为 c^+ ，负样本表示为 c^- ，当前样本全局表示为 c_a ，则对比损失写为：

$$\mathcal{L}_a^{\text{con}} = -\log \frac{\exp(\text{sim}(c_a, c^+)/\tau)}{\exp(\text{sim}(c_a, c^+)/\tau) + \sum_{c^-} \exp(\text{sim}(c_a, c^-)/\tau)} \quad (4-21)$$

其中 $\text{sim}(\cdot, \cdot)$ 表示余弦相似度， τ 为温度系数。最终训练目标写为：

$$\mathcal{L}_a = \mathcal{L}_a^{\text{cls}} + \lambda_a \mathcal{L}_a^{\text{con}} \quad (4-22)$$

通过上述设计，归因验证器能够同时提供整体归因有效性判断与槽位级结构反馈，从而为归因过程的监督与分析提供更细粒度的信号。

4.4.3.2 干预验证器

归因验证器用于判断模型在干预阶段中是否按照保持隐藏状态不变，改变反事实输入进行推理。干预验证器的结构如图4-3所示。

给定题目文本、已有事实、反事实输入和模型的干预推理过程。将模型的干预推理过程记为候选干预文本

设题目文本为 q ，事实观测为 o ，目标反事实输入为 x' ，候选干预文本为 i ，则验证器输入记为：

$$z_i = [q; o; x'; i] \quad (4-23)$$

在该输入上，验证器输出一个整体分数和两个槽位分数，分别记为：

$$\mathbf{p}_i, \quad p_x, \quad p_s \quad (4-24)$$

其中， $\mathbf{p}_i$ 评估模型反事实推理过程中干预行为的总体有效性； p_x 表示文本是否明确给出了正确的反事实输入； p_s 表示文本是否明确保持了与事实情境一致的隐藏状态或等价约束。

在模型结构上，设编码器输出的候选干预片段表示为：

$$H_i = [g_1, g_2, \dots, g_m] \quad (4-25)$$

其中 m 为候选干预文本的 token 数。对应“反事实输入”与“隐藏状态保持”两个槽位，本文引入查询向量 u_x 和 u_s ，并据此得到槽位聚合表示：

$$\beta_{k,t} = \frac{\exp(u_k^\top g_t)}{\sum_{j=1}^m \exp(u_k^\top g_j)}, \quad w_k = \sum_{t=1}^m \beta_{k,t} g_t, \quad k \in \{x, s\} \quad (4-26)$$

两个槽位分数与整体判别结果分别写为：

$$p_k = \sigma(W_k w_k + b_k), \quad k \in \{x, s\} \quad (4-27)$$

$$\mathbf{p}_i = \text{softmax}(W_i[c_i; w_x; w_s] + b_i) \quad (4-28)$$

干预验证器的训练样本同样按题目分组构造。对于同一道题目，围绕相同的题目文本 q 、事实观测 o 和目标输入 x' ，本文构造多条候选干预文本，并根据干预过程的正确性赋予不同层次标签。

在训练目标上，干预验证器同样采用分类监督与对比学习联合优化。设整体标签为 y_i ，两个槽位标签分别为 y_x 和 y_s ，则分类损失写为：

$$\mathcal{L}_i^{\text{cls}} = \text{CE}(\mathbf{p}_i, y_i) + \text{BCE}(p_x, y_x) + \text{BCE}(p_s, y_s) \quad (4-29)$$

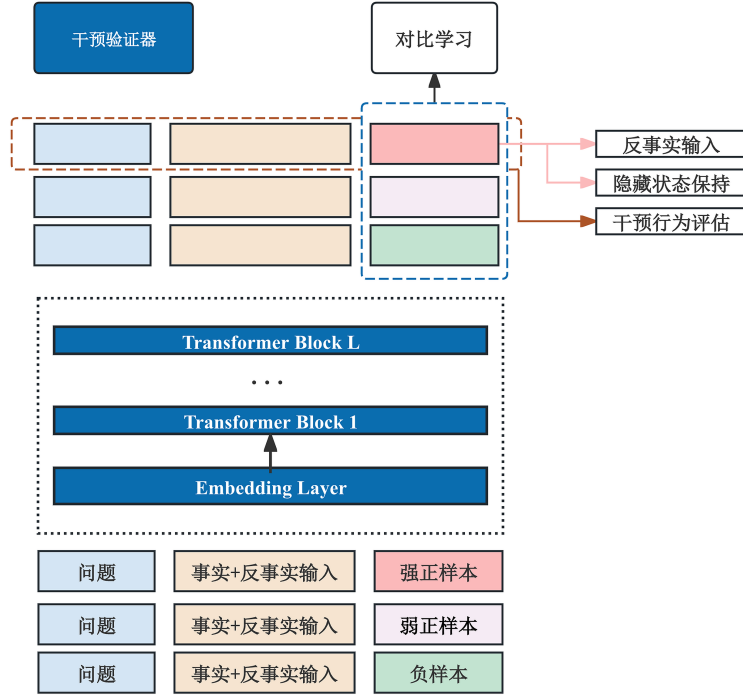


图 4-3 干预验证器结构示意图

Figure 4-3 Architecture of the intervention verifier

其中，整体标签采用类别标注，槽位标签采用二值标注。为增强同题目不同质量干预文本之间的可分性，本文在组内继续引入对比学习。设同组内强正样本表示为 c^+ ，负样本表示为 c^- ，当前样本全局表示为 c_i ，则对比损失写为：

$$\mathcal{L}_i^{\text{con}} = -\log \frac{\exp(\text{sim}(c_i, c^+)/\tau)}{\exp(\text{sim}(c_i, c^+)/\tau) + \sum_{c^-} \exp(\text{sim}(c_i, c^-)/\tau)} \quad (4-30)$$

最终训练目标写为：

$$\mathcal{L}_i = \mathcal{L}_i^{\text{cls}} + \lambda_i \mathcal{L}_i^{\text{con}} \quad (4-31)$$

通过上述设计，干预验证器能够围绕反事实输入与隐藏状态保持两个关键条件提供专门判别，为干预过程提供针对性的监督信号。

总体而言，本文通过设计归因验证器与干预验证器为反事实推理的两个关键过程：隐藏状态推断与条件替换提供监督信号。前者用于评估模型是否具备基于事实观测推断隐藏状态的能力，后者用于评估模型是否具备在共享事实基础上正确实施条件替换的能力。二者共同为强化学习训练提供过程级支撑，使模型性能变化能够进一步对应到具体过程能力的改善上。

4.5 实验

本节围绕“基于过程验证的反事实因果推理方法能否提升模型求解效果，以及这种提升是否对应关键过程行为的改善”展开实验分析。由于本文关注的是反

事实求解中的关键中间过程，实验任务不仅需要覆盖隐藏状态推断、条件替换和结果预测三个环节，还需要具备清晰的求解边界和可验证的结果基础。

基于此，本文采用代码反事实任务和数学反事实任务^[6]作为主要验证环境。这两类任务都要求模型先依据事实观测推断未显式给出的隐藏状态，再在保持该状态不变的前提下推演条件变化后的结果，因此符合反事实推理的基本形式。与此同时，这两类任务具有较清晰的求解边界，能够较稳定地分离归因、干预和预测三个环节，适合用于本文方法的有效性验证。

4.5.1 任务说明

本节进一步说明实验中使用的代码反事实任务和数学反事实任务，并介绍各任务之间的差异及其评测目标。

代码反事实任务共设置四类，分别为 `if_else`、`if_else_long`、`multi_r` 和 `while`。其中，本文在 `if_else` 任务上进行训练，并在 `if_else_long`、`multi_r` 和 `while` 三类任务上测试模型表现，以考察方法在更深条件分支、多隐藏变量和循环更新等不同程序形式下的泛化能力；此外，数学反事实任务用于进一步评估方法向自然语言场景的迁移表现。

`if_else` 任务示例如图4-4：

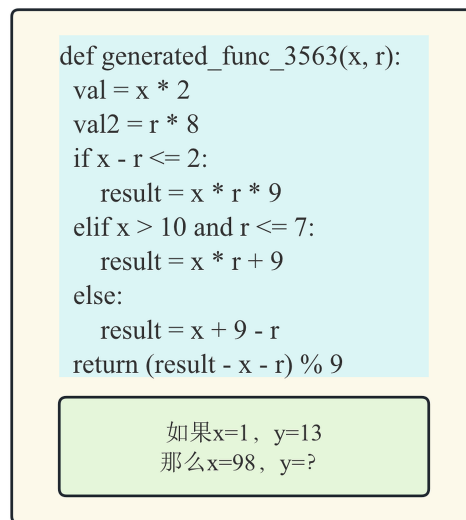


图 4-4 代码反事实任务示例

Figure 4-4 An Example of a Code Counterfactual Task

对于代码反事实任务，题目会先给出一个已知输入及其对应输出，例如“当输入 $x = 2$ 时，输出为 $y = 2$ ；如果把输入改为 $x' = 3$ ，并保持其他条件不变，输出可能是什么”。在这个任务中， x 是已知输入， r 是程序中没有直接给出的隐藏变量， y 是已知输出， x' 是改变后的输入。模型需要先根据已知的输入输出 (x, y) 推断隐藏变量 r 的可能取值，再在保持同一个 r 不变的情况下，计算输入

变为 x' 后可能得到的输出。因此，这类代码任务本质上就是一个完整的反事实推理过程。

在此基础上，`if_else_long` 包含更多层条件判断，程序结构更深，因此模型需要处理更长的推理链条。

`multi_r` 中，隐藏状态扩展为状态组 (r_1, r_2, r_3) 。因此，模型不仅要根据输入输出关系推断多个隐藏变量的可能取值，还要在保持这组隐藏状态不变的前提下完成结果推演，可用于评估多隐藏变量条件下的反事实泛化能力。

`while` 任务中，程序需要在循环过程中不断更新中间变量，并在多轮更新后得到最终输出，因此可用于考察模型在循环结构中的反事实推理能力。

数学反事实任务示例如表 4-1 所示。以该样例为例，题目显式给出每份餐盘价格、折扣比例、订购份数及已观测总价，同时保留“每份服务费”这一未直接给出的关键量。对应到前文形式化，题面中的显式条件构成事实输入 x ，隐藏的每份服务费对应隐藏状态 U ，已知总价对应事实结果 y ，将订购份数由 6 份改为 11 份对应反事实输入 x' 。模型需要先依据事实结果推理与当前总价一致的服务费取值，再在保持其余条件不变的前提下计算新的总价。由此，数学任务与代码任务在求解结构上保持一致，均可表述为围绕隐藏状态求解、条件替换和结果预测展开的反事实推理问题。

表 4-1 数学反事实任务示例

Table 4-1 A factual-counterfactual example in math tasks

场景	数学题目	答案
反事实	Ravi 正在为办公室午餐做准备。每份餐盘价格为 \$68，另有一项按每份餐盘收取的服务费。商品小计（不含服务费）可享受 14% 的折扣。订购 6 份餐盘时，页面显示总价为 \$353.88。若在其他条件保持不变的情况下改为订购 11 份餐盘，页面将显示多少金额？	\$648.78

综上，代码任务与数学任务虽然表面形式不同，但共享统一的反事实求解结构。前者用于评估模型在结构清晰环境中的反事实求解能力，后者用于评估这种能力向自然语言场景的迁移表现。基于这两类任务，本文进一步分析不同训练方式对模型反事实能力提升范围的影响。

4.5.2 数据构建

围绕上述代码反事实任务与数学反事实任务，本文基于文献^[6]提供的原始反事实样本，进一步构造四类实验数据，分别为三阶段推理数据、强化学习训练数据、归因验证器训练数据和干预验证器训练数据。原始样本同时包含题目输入与完整求解文本，其输出部分保留了从事实观测到结果预测的连续推理过程。在此基础上，本文围绕监督微调、强化学习以及过程级验证三个实验环节，对原始输出进行过程重组、标签补充与训练格式转换。

4.5.2.1 三阶段推理数据与强化学习训练数据构造

三阶段推理数据与强化学习训练数据均由原始数据集中样本转换而来。前者用于监督微调，后者用于支撑基于验证器信号的强化学习优化。围绕这一目标，本文对原始样本进行如下处理。

(1) 保留原始题目输入不变。题目中的程序文本、事实调用、事实输出以及反事实提问保持原样，以保证不同训练设置面对的是同一任务内容。

(2) 对原始求解文本进行过程切分。根据语义功能，将原始长推理中与隐藏状态推断对应的内容整理为“归因”，将与条件替换对应的内容整理为“干预”，将与替代结果预测及结果给出对应的内容整理为“预测”。

(3) 对切分后的内容进行去噪与压缩。删除原始文本中重复检查、回溯修正、姿态性表述以及与关键过程无直接关系的冗余句段，仅保留支撑归因、干预和预测所必需的核心步骤。

(4) 对每一阶段补充统一标识，并在阶段末尾加入简要结论句，使重组后的输出具有稳定边界，便于模型学习各阶段职责，也便于后续奖励计算与过程分析。

按照上述流程，可由原始自由展开式推理样本得到三阶段监督数据。其输入侧与原始样本保持一致，输出侧则重写为包含“归因”“干预”“预测”三个阶段的过程文本，数据示例如表4-2。

表 4-2 原始自由推理样本与三阶段推理样本示例

Table 4-2 Example of original free-form reasoning and three-stage reasoning sample

样本形式	内容示例
原始自由推理样本	“First, I think I need to figure out what value of r was used when x was 5. ... Wait, but the problem says ‘keeping everything else unchanged,’ so I think that means r remains the same as when x was 5. ... Wait, but we don’t know r ’s exact value, only that $r \geq 5$. So we need to consider two cases: $r \geq 15$ and $r < 15$ Wait, but let me double-check. ... So yes, y can be 1, 2, 3, 4, 5, 6, 8, 9, 11, 12.”
三阶段推理样本	“归因：For the factual call $x = 5$, ... Therefore the hidden state must satisfy $r \in \{5, 6, \dots, 600\}$. 干预：Now intervene by changing only the input from $x = 5$ to $x = 15$, while keeping the same hidden state r . 预测：For $x = 15$, the branch taken depends on whether $15 \leq r$ Therefore $y = \boxed{1, 2, 3, 4, 5, 6, 8, 9, 11, 12}$.”

由此得到的三阶段推理数据可直接用于监督微调训练。进一步地，强化学习训练数据由同一批原始样本转换得到。

除三阶段推理数据和强化学习训练数据外，本文方法还需构造两类过程级判别数据，分别用于训练归因验证器与干预验证器。

4.5.2.2 归因验证器与干预验证器训练数据构造

两个验证器的训练数据均由原始监督样本派生得到，其构造方式是在统一题目语境下围绕关键过程文本建立分组判别样本。具体而言，本文首先固定同一道题目的题目文本与事实观测，然后分别围绕归因文本和干预文本构造多条质量不同的候选文本，并为其赋予整体标签与槽位标签，从而形成适合分类监督与组内对比学习的训练样本。

归因验证器训练数据的标注规则如下：

(1) **事实观测**：若文本明确使用题目中的事实输入或事实输出，则标记为 1；否则标记为 0。

(2) **隐藏状态提及**：若文本明确提及题目中的隐藏变量、隐藏条件或未观测状态，则标记为 1；否则标记为 0。

(3) **隐藏状态推断**：若文本根据事实观测对隐藏状态进行约束、筛选、排除或范围推导，则标记为 1；否则标记为 0。

(4) **整体标签**：若文本同时满足事实观测、隐藏状态提及和隐藏状态推断三个条件，则记为强正样本；若文本包含隐藏状态但不满足强正样本条件，则记为弱正样本；若文本未提及隐藏状态，则记为负样本。

归因验证器训练数据的典型样例如表 4-3 所示。

表 4-3 归因验证器标注样例

Table 4-3 Annotation examples for the abduction verifier

样本类型	文本片段示例	标签
强正样本	“Given $x = 5$ and $y = 4$, substituting $x = 5$ into the program shows that $r < 5$ cannot yield $y = 4$, so $5 \leq r \leq 600$.”	事实观测 =1；隐藏状态提及 =1；隐藏状态推断 =1
弱正样本	“For $x = 5$ and $y = 4$, I should check how r affects the branch.”	事实观测 =1；隐藏状态提及 =1；隐藏状态推断 =0
负样本	“when $x = 5$, $y = 4$. so maybe ...”	事实观测 =1；隐藏状态提及 =0；隐藏状态推断 =0

干预验证器训练数据的标注规则如下：

(1) **反事实输入**：若文本正确设置题目要求的反事实输入，则标记为 1；若反事实输入缺失或设置错误，则标记为 0。

(2) **隐藏状态保持**：若文本正确保持事实情境中的隐藏状态不变，则标记为 1；错误改变隐藏状态，则标记为 0。

(3) **整体标签**：若文本同时满足反事实输入和隐藏状态保持两个条件，则记为强正样本；若文本正确设置反事实输入，但未明确说明隐藏状态保持，则记为弱正样本；若文本未正确设置反事实输入，或错误改变隐藏状态，则记为负样本。

表 4-4 干预验证器标注样例

Table 4-4 Annotation examples for the intervention verifier

样本类型	文本片段示例	标签
强正样本	“Change x from 5 to 15, using the same $r = 20$ as before.”	反事实输入 =1；隐藏状态保持 =1
弱正样本	“Change x from 5 to 15.”	反事实输入 =1；隐藏状态保持 =0
负样本	“Change x from 5 to 15, using $r = 9$ instead.”	反事实输入 =1；隐藏状态保持 =0

干预验证器训练数据的典型样例如表4-3。

通过上述标注规则与样例构造，归因验证器和干预验证器均能够在统一题目语境下接收层次化监督：整体标签用于学习过程文本的整体有效性，槽位标签用于学习关键局部成分是否成立，同题目下的多样化候选文本则进一步支持组内对比学习。由此得到的两类验证器训练数据可直接用于归因验证器与干预验证器的训练，并为后续强化学习提供归因信号和干预信号。

4.5.3 实验设置和训练超参数

实验基于 Qwen2.5-1.5B-Instruct^[98] 进行，并在其基础上构造如下训练设置：

- (1) Base： 未经过本章任务专门训练的基础模型。
- (2) Orig-SFT： 使用原始自由展开式推理文本进行监督微调。
- (3) Abd-SFT： 仅使用归因阶段文本进行监督微调。
- (4) Stage-SFT： 使用归因、干预和预测三阶段推理文本进行监督微调。
- (5) Origin-RL： 仅使用答案正确性奖励进行强化学习优化。
- (6) Abd-only-RL： 在答案奖励基础上加入归因验证器信号进行强化学习优化。
- (7) Abd+Int-RL： 在答案奖励和归因验证器信号基础上进一步加入干预验证器信号进行强化学习优化。

归因验证器与干预验证器单独训练，骨干模型采用预训练 BERT^[99]。

原始监督微调与三阶段推理监督微调采用相同的基础配置，学习率设为 5×10^{-5} ，训练 2 轮，最大输入长度设为 16384。基于验证器信号的强化学习学习率设为 1×10^{-6} ，训练 1 轮，最大训练步数设为 1500，输入输出长度分别设为 prompt 512 和 completion 2000，每次采样候选数设为 24，KL 系数设为 1×10^{-3} 。奖励权重设置为：答案 0.55，归因 0.20，干预 0.15，格式 0.10，截断惩罚 0.05。

归因验证器与干预验证器采用统一训练配置，学习率设为 5×10^{-5} ，训练 2 轮。

4.5.4 实验结果与分析

本节汇报前述实验设置在代码反事实任务和数学反事实任务上的实验结果，并围绕监督微调和强化学习分别开展对比实验与消融实验分析。

表 4-5 反事实任务上的实验结果

Table 4-5 Experimental results on counterfactual tasks

模型	代码反事实任务								数学反事实任务
	if_else		if_else_long		multi_r		while		GSM-CF
	F1	EM	F1	EM	F1	EM	F1	EM	Acc
Base	16.3	5.0	27.0	14.32	9.57	6.96	3.51	1.67	7.00
Orig-SFT	61.35	44.11	51.73	33.85	21.5	20.3	5.43	4.17	2.57
Stage-SFT (ours)	58.9	41.6	54.6	36.9	23.8	22.6	5.9	4.6	2.34
Origin-RL	34.7	20.2	50.3	39.6	25.5	25.2	5.0	4.2	6.38
Abd+Int-RL (ours)	36.9	21.4	54.7	42.8	31.6	31.0	8.1	6.7	12.94

注：上半部分对应监督微调对比，下半部分对应强化学习对比。GSM-CF 表示基于 GSM 风格题目构造的反事实数学任务，并非标准 GSM8K 原始测试集。

4.5.4.1 对比实验分析

表4-5展示了本文的方法与基线方法^[6]在代码反事实任务和数学反事实任务上的实验结果。对于代码反事实任务，由于同一问题可能对应多个合法答案，本文采用 F1 与精准匹配指标 EM (Exact Match) 进行评估；对于数学反事实任务，仅采用最终答案准确率 Acc 进行评估。

从表 4-5 可以看出，监督微调与强化学习在训练内代码任务、训练外代码任务和数学反事实任务上呈现出不同的表现模式。经过监督微调，Orig-SFT 在训练内 if_else 任务上提升最为明显，EM 由 Base 的 5.0 提升至 44.11，说明基于代码反事实样本的自由展开式监督能够有效增强训练分布内的求解能力。与此同时，Stage-SFT 在 if_else_long、multi_r 和 while 上整体优于 Orig-SFT，分别取得 36.9、22.6 和 4.6 的 EM，说明当任务结构发生扩展、隐藏状态复杂度提高或控制流变为循环时，更清晰的过程组织方式能够为模型提供更稳定的求解支撑。结合训练内与训练外结果可以看出，三阶段推理组织的优势主要体现在对完整反事实过程边界的显式组织，这一组织方式有助于减弱原始自由展开式长推理中过程职责交织所带来的干扰。

在领域外反事实推理任务上，结合归因验证与干预验证信号的强化学习方法表现出更明显的性能优势。Abd+Int-RL 在 if_else_long、multi_r、while 和 GSM-CF 上均优于 Base 和 Origin-RL，分别达到 42.8、31.0、6.7 和 12.94。与 Origin-RL 相比，Abd+Int-RL 在 multi_r 与 GSM-CF 上的提升尤为明显，这说明仅依赖结果层奖励虽然能够在部分训练外任务上带来基础改进，但其增益更多停留在输出层面；当奖励信号进一步覆盖隐藏状态推断与条件替换两个关键中间过程时，模

型更容易形成可迁移的反事实求解行为。换言之，强化学习的主要优势并不体现在训练内任务上的拟合，而体现在复杂结构任务和跨领域任务中的稳定泛化。

监督微调更擅长提升训练内反事实推理任务表现，而强化学习更擅长提升领域外反事实推理的泛化能力。进一步比较两条完整方法可以发现，Stage-SFT 在训练内代码任务上的表现高于 Abd+Int-RL，而 Abd+Int-RL 在训练外代码任务和数学反事实任务上的表现更优。这一差异表明，SFT 和 RL 作用于反事实求解能力的不同层面：三阶段推理组织更直接地提升模型对既有训练分布内过程结构的学习效率，验证器信号强化学习更有利于将隐藏状态推断与条件替换的约束迁移到新的程序结构和自然语言场景中。由此可以看出，前者更突出训练内求解能力增强，后者更突出结构外泛化能力增强。

图 4-5 对监督微调和强化学习微调在三类任务中的相对收益分布进行了汇总展示。图 4-5 将结果汇总为训练内代码任务、训练外代码任务和数学反事实任务三类，并分别展示监督微调与强化学习对模型的反事实因果推理能力的影响。其中，训练内代码任务使用 if_else 的 EM 指标，训练外代码任务使用 if_else_long、multi_r 和 while 的 EM 平均值，数学反事实任务使用 GSM-CF 的准确率。图中可以看到，Stage-SFT 在训练外代码任务上保持较高表现，而 Abd+Int-RL 在训练外代码任务和数学反事实任务上的优势更为突出。这一结果与表 4-5 中的逐项比较保持一致，进一步说明 SFT 和 RL 在反事实推理能力改善侧重上存在明显差异。

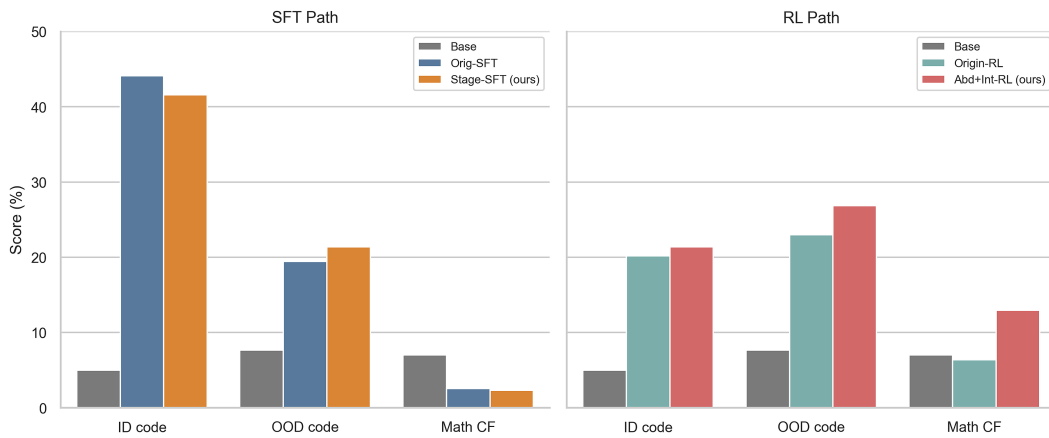


图 4-5 监督微调与强化学习在三类任务上的总体表现

Figure 4-5 Overall performance of the SFT and RL across three task groups

4.5.4.2 消融实验分析

表 4-6 展示了不同监督与奖励信号组合下的消融实验结果。监督微调的消融结果首先说明了完整推理链监督的重要性。Orig-SFT 与 Stage-SFT 的训练文本都覆盖隐藏状态推断、条件替换和结果预测三个环节，因此二者在完整反事实任务上的表现均显著优于 Abd-SFT。与之相比，Abd-SFT 只保留归因阶段，训练文本中缺失了干预与预测两部分推理链，模型虽然能够学习事实观测与隐藏状态之间的对应关系，但对条件替换后的结果预测缺少直接训练信号。因此，当测试任

表 4-6 消融实验结果
Table 4-6 Ablation results

模型	代码反事实任务								数学反事实任务
	if_else		if_else_long		multi_r		while		GSM-CF
	F1	EM	F1	EM	F1	EM	F1	EM	Acc
Orig-SFT	61.35	44.11	51.73	33.85	21.5	20.3	5.43	4.17	2.57
Abd-SFT (ours)	23.7	17.2	33.5	19.1	11.6	8.9	2.9	2.0	1.8
Stage-SFT (ours)	58.9	41.6	54.6	36.9	23.8	22.6	5.9	4.6	2.34
Origin-RL	34.7	20.2	50.3	39.6	25.5	25.2	5.0	4.2	6.38
Abd-only-RL (ours)	30.2	16.8	45.6	34.2	23.0	22.1	4.2	3.3	9.51
Abd+Int-RL (ours)	36.9	21.4	54.7	42.8	31.6	31.0	8.1	6.7	12.94

注：上半部分对应监督微调消融，下半部分对应强化学习消融。GSM-CF 表示基于 GSM 风格题目构造的反事实数学任务。

务要求输出完整反事实解答时，Abd-SFT 需要自行补足后续推理步骤，这使其在训练内和训练外代码任务上都受到明显限制。由此可以看出，在监督微调中，完整推理链的保留是支撑模型完成反事实解答的基础条件。

在完整推理链得到保留的前提下，Orig-SFT 与 Stage-SFT 的差异进一步反映了过程组织方式的作用。Orig-SFT 所使用的训练样本来自 if_else 代码反事实任务，训练文本保留了自由展开式长推理，因此其在训练内 if_else 任务上的表现较高，这与训练数据和测试任务在领域形式与推理目标上的局部对齐是一致的。Stage-SFT 在此基础上进一步将完整反事实过程组织为归因、干预和预测三个阶段，使各阶段职责更加清晰。对应到结果上，二者在训练内任务上的差距较小，而 Stage-SFT 在 if_else_long、multi_r 和 while 上整体更稳，说明三阶段推理组织的主要作用体现在训练外代码任务中。对于 if_else_long 这类分支深度增加的任务，清晰的阶段边界有助于模型维持更稳定的链式推理；对于 multi_r 和 while 这类分别对应联合搜索空间扩大和状态更新链条变长的任务，完整且边界清晰的过程组织同样更有利于模型维持求解稳定性。

强化学习的消融结果反映了不同奖励信号各自对应的过程作用。仅使用答案正确性奖励的 Origin-RL 在 if_else_long、multi_r 和 GSM-CF 上已经带来一定提升，说明结果层奖励可以为模型提供基础优化信号，并在一定程度上推动输出朝向正确结果分布调整。与此同时，该模型在 while 等任务上的提升仍然有限。这表明仅有最终答案奖励时，模型对隐藏状态推断是否充分、条件替换是否一致以及结果预测是否稳定缺少直接约束。由此可以看出，在强化学习中，答案奖励能够提供基础驱动，但尚不足以稳定覆盖完整反事实过程中的关键中间行为。

归因信号主要改善的是隐藏状态推断过程，而这一局部过程的优化未必立即带来完整反事实求解性能的提升。在答案奖励基础上加入归因验证器信号后，Abd-only-RL 在 GSM-CF 上出现进一步提升，而在代码任务上的结果未出

现同步上升。这一现象说明，归因信号确实能够增强模型对隐藏状态推断过程的关注，但奖励目标的加入也会改变原有优化重点，使模型更多朝向归因过程本身进行调整，而这一变化未必立即转化为代码反事实任务上的整体增益。对 GSM-CF 而言，数学反事实任务的主要瓶颈之一在于根据事实结果求解题目中未显式给出的未知量，因此归因信号的收益更容易显现。进一步同时加入归因验证器与干预验证器信号后，Abd+Int-RL 在 `if_else_long`、`multi_r`、`while` 和 GSM-CF 上均取得最佳结果，说明在归因信号之外，干预信号对完整反事实过程同样具有关键作用。对于 `multi_r` 这类联合搜索空间更大的任务，归因与干预信号的联合使用有助于模型更稳定地处理潜在状态筛选与条件替换；对于 `while` 这类状态更新链条更长的任务，完整的过程级奖励则更有利于维持替代情境中的多步推理一致性。由此可以看出，强化学习中的增益来源于答案奖励、归因信号和干预信号的逐步叠加，其中干预模块的引入对完整反事实求解能力的形成具有直接作用。

4.6 本章小结

本章围绕反事实因果推理任务，提出了基于过程验证的反事实因果推理方法。针对大模型在反事实因果推理中关键推理步骤缺失、条件约束信息遗漏的问题，本文从监督微调和强化学习两个方面进行方法设计。监督微调通过构建覆盖反事实完整推理过程的训练数据，使模型学习从事实观测到推导反事实结果的完整推理链；强化学习通过设计归因验证器和干预验证器，分别对反事实推理过程中隐藏状态推断和反事实条件替换两个过程进行验证，并将验证结果作为过程级奖励信号，引导模型在优化最终答案的同时关注关键中间推理过程。

实验结果表明，该方法可提升模型在反事实因果推理任务中的准确性与稳定性。监督微调实验显示，完整推理链的保留是模型完成反事实解答的基础。强化学习实验显示，答案奖励能够提供结果层面的基础优化信号，但仅依赖最终答案奖励难以充分约束中间推理过程；引入归因验证器与干预验证器后，模型在训练外代码任务和数学反事实任务上的表现进一步提升，说明过程验证能够有效缓解模型反事实推理过程中关键推理步骤缺失和条件约束信息遗漏的问题。整体而言，本章工作验证了构建细粒度信号对模型推理过程进行监督，可以有效的提升大语言模型的反事实因果推理能力。

第5章 基于因果推理的工业异常诊断系统框架

5.1 应用背景

工业生产过程通常具有工序链条长、设备类型多、工艺参数复杂等特点。工业生产过程中的异常现象往往涉及设备状态、工艺参数、环境扰动和人为操作等多类因素，其形成过程通常跨越多个工序节点。与此同时，支撑分析的证据又分散于制造执行系统、设备监测系统、质量管理体系、工艺文档和历史案例库等不同载体之中。这使得工业异常诊断需要同时完成异常信息汇聚、异常成因识别、传播链解释和处置决策等多类分析任务。因此，面向工业生产异常诊断场景设计具备多源信息整合能力的自动化异常诊断框架，具有明确的现实需求。

从现有生产实践看，工业异常诊断工作仍较多依赖人工排查和专家会诊。传统处理方式通常以异常工单、现场经验和局部数据分析为主要依据，在复杂异常场景中容易受到多源数据分散、分析过程依赖个人经验以及异常传播链路难以识别等因素影响。如何将分散证据组织为可比较、可追踪、可复用的诊断依据，是工业异常诊断系统框架设计需要解决的重要问题。

本文前两个研究点的内容可以为工业异常诊断中的根因分析和改进生成提供方法支撑。基于第三章的研究内容，本文构建了溯因分析模块，并用于工业异常诊断场景中分析造成当前生产异常的关键因素及其作用过程。该模块将异常现象、相关工序、过程事件和证据来源组织为局部工业事理图谱，并通过因果关系补全分析识别异常形成过程中的潜在传播路径，为复杂异常场景下的根因收敛提供支撑。基于第四章的相关研究，本文构建了改进生成模块，并用于工业异常诊断场景中生成解决当前生产异常、降低后续质量风险的处置建议。该模块将根因识别后的处置分析组织为反事实推理任务，在保持关键生产约束基本一致的条件下比较不同控制措施对异常结果、质量风险和生产连续性的影响，为现场处置和后续预防提供改进建议。

基于此，本研究面向工业生产异常诊断场景，构建了基于因果推理的工业异常诊断系统框架。在实现上，多智能体协同、检索增强生成等技术为前文方法嵌入工业诊断流程提供了技术支撑。该框架能够在异常发生后自动汇聚相关证据，分析异常产生的主要原因，并推演不同处置措施可能带来的影响，用于工业生产过程故障的自动化诊断与处置决策支持。

5.2 应用整体框架

本节内容包括整体业务流程、总体架构、数据分析模块、溯因分析模块、改进生成模块、数据支撑模块的介绍。详细介绍了工业异常诊断的流程和方法，以及相关模块的细节。下面对各部分内容进行详细说明。

5.2.1 整体业务流程

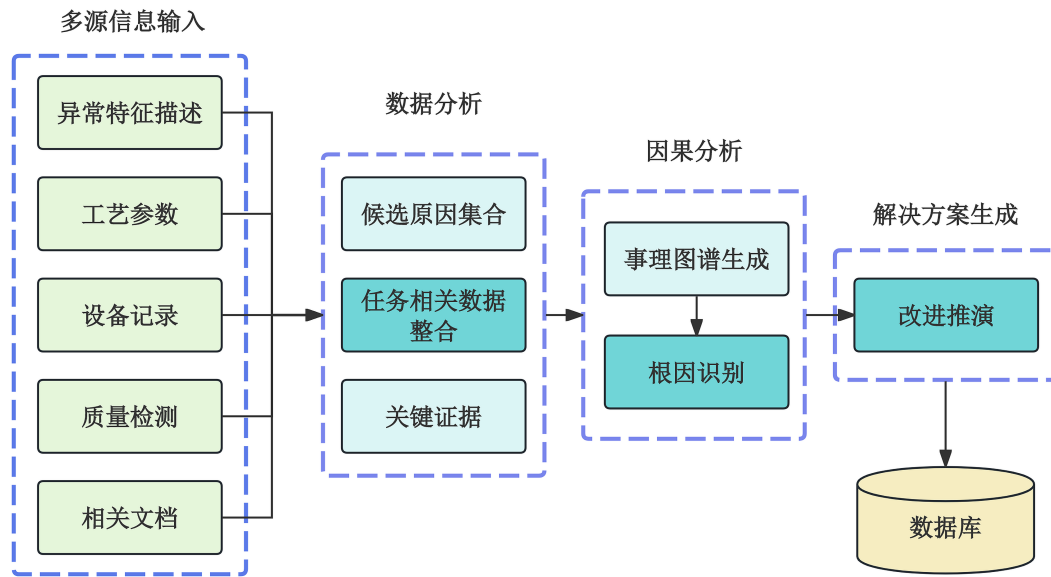


图 5-1 整体业务流程

Figure 5-1 Overall workflow

整体业务流程如图 5-1 所示。当生产现场出现质量偏差、设备异常、工艺波动或安全隐患时，异常工单首先被提交至诊断入口，并同步关联设备状态、工艺参数、质量检测结果、工艺文档和历史案例等信息。系统框架根据异常对象、发生工序、时间窗口和关联设备提取相关数据，将分散在制造执行系统、设备监测系统、质量管理系统和知识库中的信息整理为统一任务上下文。在任务上下文形成后，相关工艺记录、设备状态曲线、检测偏差、报警记录和相似案例会被进一步汇聚，形成初始候选原因和关键证据。随后，系统框架围绕异常结果构建局部工业事理图谱，将候选原因、工序节点、设备状态、工艺参数、质量表现和证据来源组织到同一图结构中，并结合知识图谱关系和因果分析模型补全候选原因与异常结果之间可能存在的因果联系，得到候选传播路径、关键证据和主导根因判断。在根因识别完成后，系统框架根据主导根因、异常传播链和现场约束，对工艺参数调整、设备状态恢复、工装维护和复检范围调整等处置方案进行推演分析，形成处置建议、优化建议和预防建议。最终，诊断结论、异常传播链、改进建议和分析报告返回现场处置环节，并同步回写至历史案例库和知识图谱，用于后续相似异常诊断和知识复用。

5.2.2 总体架构

整体系统架构如图 5-2 所示，该框架分为模块层、接口层和应用层。模块层包括数据分析模块、溯因分析模块、改进生成模块和数据支撑模块，是本章后续重点介绍的部分；接口层包括通用工具和服务。应用层主要负责稳态监控、趋势



图 5-2 总体架构

Figure 5-2 Overall Architecture

分析、工单管理和智能对话等业务功能。本章只对核心模块进行介绍。

5.2.2.1 数据分析模块

数据分析模块面向工业异常诊断中的多源信息汇聚问题，主要完成异常任务接入、证据召回、证据组织。该模块采用智能体协同编排方式实现异常工单解析、检索工具调用、证据模板填充和任务包生成，并通过 AgentScope 框架完成角色配置、消息传递和工具调用记录。工业生产现场出现异常告警后，系统触发数据分析流程，由编排智能体维护任务状态并分配子任务，数据分析智能体、知识检索智能体和领域复核智能体围绕同一异常任务协同完成信息汇聚。

在具体执行过程中，编排智能体首先根据异常工单中的产品编号、批次号、工序编号、设备编号、异常类型和发生时间，确定本次诊断任务的对象边界、工序范围和时间窗口。随后调用知识检索智能体和数据分析智能体，围绕异常对象、关联工序、关键设备和发生时间开展混合检索，通过 BM25 关键词检索、向量检索和结构化条件查询召回与异常相关的生产过程数据、设备状态数据、质量检测数据、工艺知识和历史案例信息。

在召回与异常诊断相关的数据后，数据分析智能体按照预设的诊断信息模板对多源数据进行整理，并对不同来源证据进行字段对齐、时间对齐和来源标

注。对于生产过程数据，模块重点提取工艺参数设定值、实际执行值、关键工序时间和批次流转关系；对于设备状态数据，模块重点提取报警记录、状态曲线、点检记录和维护记录；对于质量检测数据，模块重点提取不合格项、检测偏差、复检结果和异常件分布；对于知识与案例数据，模块重点提取工艺约束、故障模式、图谱关系和相似案例。领域复核智能体进一步检查证据是否覆盖异常对象、相关工序、关键设备和质量表现，并判断候选原因与工艺机理、设备状态和质量表现之间是否一致。

在智能体编排流程中，编排智能体负责维护数据分析任务状态并判断是否需要补充检索；数据分析智能体负责从过程数据、设备数据和质量数据中提取异常波动、统计摘要和关键证据；知识检索智能体负责补充工艺规范、历史案例和知识图谱关系；领域复核智能体负责对证据完整性、候选原因合理性和跨源信息一致性进行检查。当存在数据缺失或证据冲突时，编排智能体会根据补充信息清单再次触发检索或复核。

通过数据分析模块，在进入溯因分析之前会形成一个诊断任务包。该任务包包括异常任务上下文、多源证据摘要、初始候选原因集合和补充信息清单。其中，异常任务上下文用于明确诊断边界，多源证据摘要用于记录各类证据及其来源，初始候选原因集合用于给出后续因果比较的候选对象，补充信息清单用于标记仍需复核的数据缺失或证据冲突。

5.2.2.2 溯因分析模块

溯因分析模块面向工业异常诊断中的根因识别任务，主要解决候选原因多、证据链不完整和跨工序传播关系难以判断的问题。经过数据分析模块处理后，诊断任务已经形成异常任务上下文、多源证据摘要和初始候选原因集合。溯因分析模块在此基础上进一步构建面向异常结果的因果分析结构，对候选原因进行传播路径补全、贡献度比较和根因排序。该模块主要由工业事理图谱构建、因果关系补全分析和根因校验三个部分组成，如表 5-1 所示。

表 5-1 溯因分析模块组成

Table 5-1 Components of the Abductive Analysis Module

组成部分	主要输入	主要输出
工业事理图谱构建	异常任务上下文、多源证据摘要、初始候选原因、知识图谱关系	局部工业事理图谱、候选原因节点、异常结果节点、证据节点
因果关系补全分析	局部工业事理图谱、候选原因节点、异常结果节点	补全后的因果边、候选传播路径、候选原因因果强度
根因校验	候选传播路径、因果强度、关键证据、领域约束	候选原因排序、主导根因判断、异常传播链、待复核因素

工业事理图谱构建。这部分面向异常诊断中的证据结构化组织问题，主要完成异常对象、候选原因、过程事件和证据来源之间的图结构表达。该部分采用图

谱构建智能体实现,智能体读取数据分析模块输出的异常任务上下文、多源证据摘要和初始候选原因集合,并调用数据支撑模块中的工艺路线、设备台账、质量规则和知识图谱关系。具体执行时,模块首先根据异常任务上下文确定异常结果节点和关联工序节点;随后从候选原因集合与多源证据摘要中提取原因节点、过程节点和证据节点;最后依据工序顺序、设备关联、参数影响、状态变化、质量传递和证据指向等关系生成边,并结合已有知识关系对局部图结构进行补充,形成面向当前异常任务的局部工业事理图谱,为后续因果关系补全和根因排序提供结构化输入。

因果关系补全分析。这部分用于判断候选原因与异常结果之间是否存在可解释的传播链路。该部分通过因果归纳边补全模型实现,模型以局部工业事理图谱、候选原因节点和异常结果节点为输入,在图结构上下文中预测候选原因指向异常结果的因果强度。对于因果强度较高且符合领域约束的候选关系,模块将其作为补全因果边加入局部图谱,并进一步搜索候选原因到异常结果之间的传播路径。

根因校验。这部分用于在候选传播路径之间形成根因判断。该部分通过根因校验智能体实现,智能体读取因果关系补全分析输出的候选传播路径、候选原因因果强度和关键证据,并结合数据支撑模块中的工艺约束、设备机理和质量判定规则进行综合评估。具体执行时,模块首先按照因果强度对候选原因进行初步排序;随后检查每条候选路径是否具有完整的中间过程、是否得到多源证据支撑、是否符合领域约束;最后将能够同时解释主要异常现象且证据链较完整的因素识别为高可信候选根因,将证据不足或与领域约束冲突的因素标记为待复核对象。通过上述过程,模块输出候选原因排序、主导根因判断、关键证据和异常传播链。

通过上述三个部分,溯因分析模块将数据分析模块输出的任务包进一步转化为根因诊断结果。其最终输出包括候选原因排序、关键证据、补全后的异常传播链、主导根因判断和待复核因素。这些结果一方面为改进生成模块提供主导根因、关键控制因素和异常传播链路,另一方面也为结果展示、案例回写和知识沉淀提供可追踪的诊断依据。

5.2.2.3 改进生成模块

改进生成模块面向根因识别之后的处置推演任务,用于将根因诊断结果转化为现场可执行的生产改进方案。该模块首先读取主导根因、异常传播链、关键控制因素和现场约束,并将改进生成过程组织为反事实推演任务。具体而言,任务输入包括异常现状、拟调整的控制条件和需保持不变的生产约束,其中异常现状由当前批次、工序路线、设备状态、工艺参数和质量规则构成,拟调整的控制条件对应参数校正、设备维护、工装更换、复检范围调整等候选措施,需保持不变的生产约束用于限定推演边界。随后,模块调用经过过程验证训练的改进生成模型,分析不同候选措施实施后异常结果、质量风险和生产连续性可能发生的变

化，并据此形成处置建议、优化建议和预防建议。

5.2.2.4 数据支撑模块

数据支撑模块用于统一管理生产过程数据、设备状态数据、质量检测数据、工艺文档、知识图谱、历史案例库和诊断结果库。其中，生产过程数据和设备状态数据用于支撑异常任务上下文构建和候选原因分析，质量检测数据用于刻画异常现象和验证诊断结果，知识图谱和历史案例用于支撑图谱补充、因果路径分析和建议复用。

每次诊断完成后，候选原因集合、关键证据、异常传播链、主导根因、改进建议和分析报告会被回写至该模块，并根据需要更新知识图谱中的事件关系和处置策略。

5.3 应用案例分析

为说明本文构建的系统框架在复杂工业生产异常诊断中的有效性，本节选取航空发动机涡轮盘热处理与精加工联合生产过程中的异常诊断任务进行分析。

5.3.1 案例背景与问题描述

该任务来自某航空零部件制造场景中的涡轮盘生产过程。该零件在完成固溶时效热处理后进入精车和磨削工序，最终需满足尺寸精度、表面粗糙度和硬度一致性等质量要求。在一批次生产过程中，质检环节连续发现多件零件存在内孔尺寸超差与局部表面烧伤并发的的问题，且异常并未稳定集中于单一机台，导致现场难以通过经验快速定位原因。

该异常任务的关键信息如表 5-2 所示。

表 5-2 异常任务背景信息
Table 5-2 Background Information of the Task

项目	内容
产品对象	航空发动机涡轮盘
异常批次	2025 年 4 月某批次连续生产件，共 26 件，其中 7 件异常
主要工序	固溶处理、时效处理、精车、磨削、终检
异常现象	内孔尺寸超出上公差 0.018–0.034 mm，同时出现局部表面烧伤痕迹
影响范围	涉及 2 台数控车床、1 台内圆磨床和对应热处理批次，异常跨工序传播特征明显
已知约束	原材料批次一致，操作班组相同，检测设备经校准后未发现明显失准

从任务特征看，其诊断难点主要体现在三个方面：第一，异常同时表现为尺寸偏差和表面烧伤，说明问题可能跨越热处理质量、切削稳定性和磨削状态等多个环节；第二，异常件并未全部集中于单一设备，说明需要结合工序流转、设备

状态和质量表现分析异常的跨工序传播关系；第三，热处理炉温波动、磨削主轴电流升高、冷却液压力变化和历史相似案例等证据分散在不同数据源中，使得单一数据源难以支撑完整诊断判断，增加了跨源证据关联、候选原因比较和传播链识别的难度。

5.3.2 异常诊断过程分析

诊断过程首先由数据分析模块围绕异常工单形成统一任务上下文。编排智能体根据产品对象、异常批次、主要工序、关联设备、异常类型和发生时间确定检索边界，并调用数据分析智能体和知识检索智能体，通过关键词检索、向量检索和结构化条件查询召回热处理炉温、磨床主轴电流、冷却液压力、砂轮修整记录、工装点检记录、终检结果、工艺规范和历史案例等信息。随后，数据分析智能体按照诊断信息模板整理多源证据，领域复核智能体检查证据覆盖范围和候选原因合理性，最终形成诊断任务包。该阶段形成的诊断任务包如表 5-3 所示。

表 5-3 异常诊断任务包
Table 5-3 Diagnostic Task Package for the Anomaly Task

数据类别	召回数据与关联证据	任务包整理结果
生产过程数据	固溶末段炉温曲线、磨削时间、砂轮修整记录、工序流转记录	炉温波动、磨削时间变化、砂轮修整异常、工序流转关系
设备状态数据	磨床主轴电流、冷却液压力、设备报警和点检记录	主轴电流变化、冷却液压力变化、设备状态异常、点检异常记录
质量检测数据	内孔尺寸检测、表面烧伤记录、硬度复检结果、异常件分布	尺寸偏差、表面质量异常、硬度偏差、异常件分布特征
知识与案例数据	工艺规范、质量规则、知识图谱关系、历史相似案例	工艺约束、质量判定规则、候选原因线索、相似案例证据

基于上述任务包，数据分析模块将工艺、设备、质量和历史案例证据整理为多源证据摘要，并形成热处理温度异常、磨削状态异常、工装定位偏差、检测设备漂移和原材料批次波动等初始候选原因集合。

进入溯因分析模块后，图谱构建智能体读取诊断任务包，将异常结果、关联工序、候选原因、过程事件和证据来源组织为局部工业事理图谱。在该图谱中，热处理炉温上冲、硬度偏高、主轴电流升高、冷却液压力波动、砂轮修整间隔异常、工装定位销磨损和检测结果被组织为过程节点或证据节点，工序先后、参数影响、状态变化和质量传递关系被组织为图谱边。随后，因果关系补全分析调用已训练的因果归纳边补全模型，对候选原因节点与异常结果节点之间的潜在因果关联强度进行预测，并结合路径搜索形成候选传播链。

在根因校验阶段，根因校验智能体读取候选传播路径、候选原因因果强度和关键证据，并结合工艺约束、设备机理和质量判定规则对候选原因进行综合评估。综合炉温记录、硬度偏差、主轴电流、冷却液压力、砂轮修整记录和复检结果等证据，热处理炉温短时上冲与磨削阶段砂轮钝化叠加冷却不足被判定为高

可信候选根因。其中，前者主要解释材料局部硬化和后续加工负载上升，后者主要解释表面烧伤和局部尺寸进一步漂移。车削工装定位销磨损能够解释部分零件内孔尺寸离散加大，被保留为次要因素；检测设备漂移和原材料批次波动缺乏校准记录、复检结果和入厂检验记录支持，被标记为低可信或待复核因素。候选原因比较结果如图 5-3 所示。



图 5-3 候选原因比较结果
Figure 5-3 Comparison of Candidate Causes

5.3.3 分析结果与应用效果

综合数据分析和溯因分析结果，该异常任务的主导根因可归纳为热处理固溶末段炉温短时上冲与磨削阶段砂轮钝化叠加冷却不足。其中，热处理温度异常是上游诱因，主要导致材料局部硬度偏高和后续加工负载上升；磨削状态异常是直接触发因素，主要导致磨削热积累、局部表面烧伤和尺寸补偿失效。工装定位销磨损对部分零件内孔尺寸离散具有放大作用，但其解释范围有限，因此被识别为次要因素。

在改进生成环节，改进生成模块将处置分析组织为反事实推演任务。该任务以当前异常现状作为事实情境，包括同批次生产对象、既有工序路线、设备状态、关键工艺参数、质量检测结果和已识别的异常传播链；以热处理温控校正、砂轮修整周期缩短、冷却液压力恢复、磨削进给参数调整、工装定位销更换和复检范围扩大等措施作为候选调整条件；以原材料批次、主要工艺路线、质量判定规则和现场生产约束作为保持不变的条件。在此任务格式下，改进生成模块调用经过过程验证训练的改进生成模型，对不同候选措施实施后的结果变化进行推演。推演结果表明，温控回路校正和批次末段温度预警调整主要降低后续零件硬

度偏高风险；缩短砂轮修整周期并恢复冷却液压力主要降低磨削热积累和表面烧伤风险；更换定位销并提高工装点检频次主要改善部分零件的尺寸离散问题。根据上述推演结果，系统框架对参数校正、设备点检、修整周期调整、复检范围扩大和在制品分级处置方案进行排序，并形成表 5-4 所示的处置建议。

表 5-4 异常诊断结果与处置建议
Table 5-4 Anomaly Diagnosis Results and Recommended Actions

项目	诊断结论	建议措施
主导根因 1	热处理固溶末段炉温短时上冲，造成局部组织硬化	校正温控回路参数，增加批次末段温度预警阈值，对同炉批次追加硬度复检
主导根因 2	磨削阶段砂轮钝化叠加冷却不足，引起热积累和表面烧伤	缩短砂轮修整周期，排查冷却液压力波动原因，调整磨削进给参数
次要因素	工装定位销磨损导致部分零件装夹重复定位精度下降	更换定位销并提高工装点检频次，对相关机台进行补偿校验
处置范围	涉及同批次热处理件及关联机台后续在制品	对在制品分级复检，并暂停相同参数组合下的连续生产
知识沉淀	形成跨工序复合异常案例	将本次因果链条和处置策略写入案例库与知识图谱

在结果输出之后，诊断结论、关键证据、异常传播链和改进建议被回写至历史案例库和知识图谱，并支持后续分级处置。对于同炉批次的在制品，建议追加硬度复检与尺寸复测；对于关联磨床和车削工装，建议同步开展设备点检、修整参数调整和工装精度校验；对于后续生产批次，则通过温度预警阈值调整、砂轮修整周期优化和案例复用机制提供预防性支持。

从分析结果看，该框架的应用效果主要体现在以下三方面：

- (1) 数据分析模块将分散在生产过程、设备状态、质量检测和历史案例中的信息整理为统一诊断任务包，降低了多源证据分散对诊断过程的影响。
- (2) 溯因分析模块通过局部工业事理图谱、因果关系补全和根因校验，将候选原因、传播路径和关键证据组织为可追踪的诊断依据。
- (3) 改进生成模块将处置方案组织为反事实推演任务，比较不同候选措施对异常结果、质量风险和生产连续性的影响，形成分级处置和预防建议。

上述分析结果表明，本文构建的工业生产异常诊断系统框架能够将任务包生成、因果链补全、根因校验、改进生成和结果沉淀纳入统一业务链路，在复杂制造场景下支撑可追踪的异常诊断与处置决策。

5.4 本章小结

本章围绕工业生产异常诊断场景中的系统化应用问题，对应用背景、整体业务流程、总体架构及关键模块设计进行了说明。针对工业异常任务中异常成因复杂、证据来源分散、分析过程依赖人工经验等特点，本文构建了基于因果推理的

工业异常诊断系统框架。该框架核心模块由数据分析模块、溯因分析模块、改进生成模块和数据知识支撑模块组成，能够支撑异常任务接入、多源证据组织、根因分析、处置推演和结果沉淀。

在系统运行链路中，前文提出的两类方法被转化为可调用的分析能力。其中，第三章相关研究支撑溯因分析模块，用于分析造成当前生产异常的关键因素及其作用过程；第四章相关研究支撑改进生成模块，用于在根因识别之后推演不同处置措施可能带来的影响，并生成面向现场处置和后续预防的改进建议。通过这种组织方式，前文方法被嵌入工业异常诊断链路，形成从异常证据汇聚到处置建议生成的连续分析过程。

在应用验证方面，本章选取航空发动机涡轮盘生产过程中的复合异常任务作为分析对象。分析结果表明，所构建系统框架能够在复杂工业场景下支撑异常任务组织、因果链补全、根因校验和改进生成，并支持诊断结果回写与知识沉淀，体现出面向实际业务流程的应用价值。

总体而言，本章完成了面向工业生产异常诊断场景的系统框架设计、实现与应用说明。通过将前文方法组织为溯因分析和改进生成能力，所构建框架能够将多源证据汇聚、异常原因分析、处置措施推演和诊断知识复用纳入统一链路，为后续在更多工业生产场景中的推广应用提供基础。

第6章 总结与展望

6.1 本文工作总结

本研究针对现有大语言模型在因果推理任务上的缺陷，展开基于图谱补全和过程验证的因果推理研究与应用。并通过实验验证了本文提出的方法的有效性和实际应用价值。研究成果可归纳如下。

(1) 基于事理图谱补全的因果推理方法

首先，针对大模型在因果判断中依赖统计相关性、易受表层线索干扰的问题，本文提出基于事理图谱补全的因果推理方法。实验结果表明，该方法可提升长文本情境下因果归因的准确性与稳定性。针对长文本中背景信息冗余、候选原因竞争、时序与条件线索隐含等特点，本文将原始文本中的事件及其依赖关系组织为结构化事理图谱，并通过多次抽取一致性校验、因果归纳边补全和候选原因排序，将分散在文本中的关键结构证据进一步凝练为支持最终判断的因果摘要。实验结果表明，现有主流大语言模型在长文本情境因果判断任务上的基础表现仍然有限，提示增强方法的收益也不稳定；相比之下，基于事理图谱补全的结构化方法能够更有效地压缩冗余信息、突出关键因果路径，并在不同模型规模上取得稳定增益。相关结果说明，在复杂文本情境下，引入面向事件关系中间表示，有助于减弱模型对表层文本顺序和局部显著信息的依赖，从而提升因果判断的稳定性与有效性。

(2) 基于过程验证的反事实因果推理方法

其次，针对大模型在反事实因果推理中关键推理步骤缺失、条件约束信息遗漏的问题，提出基于过程验证的反事实因果推理方法。实验结果表明，该方法可提升模型在反事实因果推理任务中的准确性与稳定性。使用包含完整反事实推理过程的数据对模型进行监督微调，能够在保持训练内任务性能的同时提升模型在部分训练外任务上的稳定性，而结合验证器信号的强化学习方法在多隐藏变量、新控制流以及跨领域任务上表现更优。相关结果说明，围绕反事实推理关键中间过程设计细粒度的监督和验证信号，更有助于提升模型的反事实因果推理效果及其跨结构、跨任务的泛化表现。

(3) 基于因果推理的工业异常诊断系统框架

最后，面向工业生产异常诊断任务的痛点，本文设计和实现了基于因果推理的工业异常诊断系统。真实业务上的案例分析表明：前述研究能够为工业异常诊断中的异常成因识别、传播链路解释和处置建议生成提供有效支撑，这说明，以大语言模型为技术载体的因果推理方法研究不仅具有理论价值，也具备面向复杂工业场景开展应用验证和工程落地的现实价值。

总体而言，本文在以下几个方面取得了阶段性成果：一是围绕长文本情境中的因果判断问题，提出了基于事理图谱补全的因果推理方法，验证了结构化中间

表示对于复杂因果判断任务的积极作用；二是围绕反事实推理问题，提出了基于过程验证的反事实因果推理方法，验证了关键过程监督信号对于提升模型反事实推理效果的重要价值；三是围绕工业生产异常诊断场景，说明了以大语言模型为技术载体的因果推理方法在实际业务环境中的可应用性。上述研究为进一步推动大语言模型从表层语言处理走向机制性分析、从离线评测走向复杂场景应用提供了较为系统的研究基础。

6.2 未来工作展望

尽管本文围绕基于图谱补全和过程验证的因果推理研究及其工业场景应用开展了较为系统的研究，但仍存在若干有待进一步深入的问题。结合本文的研究过程与实验结果，未来工作可从以下几个方面展开。

第一，进一步提升事理图谱结构化表示的跨领域泛化能力。本文在长文本情境因果判断任务中引入事理图谱补全，有效缓解了原始文本空间中因果信息分散和表层干扰较强的问题，但当前事理图谱的关系类型体系仍与具体任务设定、人工定义的关系类别以及领域文本特征密切相关。特别是在不同文本体裁和不同业务领域中，事件粒度、关系表达方式和因果机制形态可能存在明显差异，现有结构化表示在医疗、金融、公共治理等高价值因果推理场景中的适配性仍需进一步验证。未来可进一步研究事理图谱关系类型的自动学习与动态更新机制，探索将事件结构、领域常识、常识因果知识和动态因果机制进行统一建模的方法，构建更加通用的结构化因果表示框架，从而提升方法在跨领域、跨文本体裁和开放因果任务中的泛化能力。

第二，进一步加强反事实因果推理过程的细粒度建模与验证。本文已从反事实数据构建和关键过程验证信号设计两个角度对反事实因果推理方法进行了探索，但当前过程验证主要集中于归因和干预两个核心环节，对预测阶段中间推导步骤的约束仍然不足。对于多隐藏变量强耦合、长时序依赖、高阶因果约束以及多主体交互等更复杂的反事实任务，现有归因—干预—预测三阶段划分仍有进一步细化空间。未来可以围绕隐藏状态联合推断、条件替换一致性、预测链条分步校验和结果不确定性表达等方面构建更加细粒度的监督与验证信号，并在真实业务场景中的复杂反事实任务上检验模型的过程稳定性、可迁移性和理论普适性，从而更加全面地刻画大语言模型在反事实推理中的能力边界。

第三，进一步扩大真实场景中的数据规模与应用验证范围。本文以公开数据集和工业生产异常诊断案例为基础，对所提出方法进行了实验评估和应用分析，但当前真实业务验证仍主要围绕单一工业案例展开，尚不足以充分检验方法在多行业、大规模真实数据和复杂异常类型下的鲁棒性与可扩展性。未来可在更多工业过程、更多设备类型、更多异常模式以及更长周期的真实生产数据条件下开展系统评估，重点考察方法在数据噪声、知识缺失、异常传播链复杂化和业务规则变化条件下的稳定性。同时，可进一步拓展到医疗、金融等高价值因果推理场景，围绕疾病成因分析、治疗方案反事实评估、金融风险传导和策略干预效果分

析等任务开展验证，从而更充分地评估本文方法在不同领域中的适用边界和应用价值。

综上，围绕基于图谱补全和过程验证的因果推理研究及其应用仍处于持续发展阶段。后续研究需要在结构化因果表示的自动学习与跨领域泛化、反事实推理过程的细粒度建模与验证、以及真实业务场景中的大规模应用评估等方面继续深入。随着模型能力、工具协同方式和场景应用需求的不断演进，如何使大语言模型在更复杂的因果判断、反事实因果推理和实际决策支持任务中表现出更高的稳定性、可解释性和应用价值，仍是值得持续探索的重要方向。

参考文献

- [1] Chi H, Li H, Yang W, et al. Unveiling causal reasoning in large language models: Reality or mirage? [J]. *Advances in Neural Information Processing Systems*, 2024, 37: 96640-96670.
- [2] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners [J]. *Advances in Neural Information Processing Systems*, 2020, 33: 1877-1901.
- [3] Yao S, Yu D, Zhao J, et al. Tree of thoughts: Deliberate problem solving with large language models [J]. *Advances in Neural Information Processing Systems*, 2023, 36: 11809-11822.
- [4] Jin Z, Chen Y, Leeb F, et al. Cladder: Assessing causal reasoning in language models [J]. *Advances in Neural Information Processing Systems*, 2023, 36: 31038-31065.
- [5] Li Y, Shen Y, Nian Y, et al. Mitigating hallucinations in large language models via causal reasoning [C]//*Proceedings of the AAAI Conference on Artificial Intelligence: volume 40*. 2026: 31852-31860.
- [6] Vashishtha A, Dai Q, Mei H, et al. Executable counterfactuals: Improving LLMs' causal reasoning through code [C]//*The Fourteenth International Conference on Learning Representations*. 2026.
- [7] Peters J, Janzing D, Scholkopf B. *Elements of causal inference: foundations and learning algorithms* [M]. MIT press, 2017.
- [8] Pearl J. *Causality* [M]. Cambridge University Press, 2009.
- [9] Pearl J. *The book of why: The new science of cause and effect* [M]. Basic Books, 2018.
- [10] Pearl J. Causal inference [J]. *Causality: Objectives and Assessment*, 2010: 39-58.
- [11] Imbens G W, Rubin D B. *Causal inference in statistics, social, and biomedical sciences* [M]. Cambridge University Press, 2015.
- [12] Sanchez P, Voisey J P, Xia T, et al. Causal machine learning for healthcare and precision medicine [J]. *Royal Society Open Science*, 2022, 9(8).
- [13] Cunningham S, Seward J. A brief introduction to causal inference in health care [J]. *The Cambridge Handbook of Productivity, Efficiency and Effectiveness in Health Care*, 2024.
- [14] Johnson K W, Glicksberg B S, Hodos R A, et al. Causal inference on electronic health records to assess blood pressure treatment targets: an application of the parametric g formula [C]//*Pacific Symposium on Biocomputing 2018: Proceedings of the Pacific Symposium*. World Scientific, 2018: 180-191.
- [15] Butzin-Dozier Z, Ji Y, Wang L C, et al. Causal inference via electronic health records in the national clinical cohort collaborative: Challenges and solutions in long covid research [J]. *MedRxiv*, 2025.
- [16] Komura T, Watanabe M, Shioda K. Exploring the application of target trial emulation in vaccine evaluation: scoping review [J]. *American Journal of Epidemiology*, 2025, 194(10): 3028-3040.
- [17] Wang Y, Tang J, Li C. Registration reform and stock mispricing: Causal inference based on double machine learning [J]. *Research in International Business and Finance*, 2025, 73: 102668.
- [18] Pitso S D, Michael T. Bayesian causal inference for credit default risk [J]. *Risks*, 2026, 14(2): 38.

- [19] Ogburn E L, Sofrygin O, Diaz I, et al. Causal inference for social network data [J]. *Journal of the American Statistical Association*, 2024, 119(545): 597-611.
- [20] Tay L Q, Hurlstone M, Jiang Y, et al. Causal inference in misinformation and conspiracy research [J]. *Advances in Psychology*, 2024, 2: e69941.
- [21] Park K, Kwak H, An J, et al. How-to present news on social media: A causal analysis of editing news headlines for boosting user engagement [C]//*Proceedings of the International AAAI Conference on Web and Social Media: volume 15*. 2021: 491-502.
- [22] Xiao X, Alharbi K, Zhang P, et al. Bayesian federated causal inference and its application in manufacturing [J]. *Journal of Intelligent Manufacturing*, 2025: 1-25.
- [23] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 27730-27744.
- [24] Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models [J]. *Journal of Machine Learning Research*, 2024, 25(70): 1-53.
- [25] Longpre S, Hou L, Vu T, et al. The flan collection: Designing data and methods for effective instruction tuning [C]//*International Conference on Machine Learning*. PMLR, 2023: 22631-22648.
- [26] Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: Language models can teach themselves to use tools [J]. *Advances in Neural Information Processing Systems*, 2023, 36: 68539-68551.
- [27] Egami N, Fong C J, Grimmer J, et al. How to make causal inferences using texts [J]. *Science Advances*, 2022, 8(42): eabg2652.
- [28] Veitch V, Sridhar D, Blei D. Adapting text embeddings for causal inference [C]//*Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020: 919-928.
- [29] Roberts M E, Stewart B M, Nielsen R A. Adjusting for confounding with text matching [J]. *American Journal of Political Science*, 2020, 64(4): 887-903.
- [30] Keith K, Jensen D, O' Connor B. Text and causal inference: A review of using text to remove confounding from causal estimates [C]//*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020: 5332-5344.
- [31] Keith K, Rice D, O' Connor B. Text as causal mediators: research design for causal estimates of differential treatment of social groups via language aspects [C]//*Proceedings of the First Workshop on Causal Inference and NLP*. 2021: 21-32.
- [32] Liu X, Xu P, Wu J, et al. Large language models and causal inference in collaboration: A comprehensive survey [J]. *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025: 7668-7684.
- [33] Ma J. Causal inference with large language model: A survey [J]. *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025: 5886-5898.
- [34] Li X, Cai Z, Wang S, et al. A survey on enhancing causal reasoning ability of large language models [C]//*Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2025: 399-416.
- [35] Zečević M, Willig M, Dhami D S, et al. Causal parrots: Large language models may talk causality but are not causal [J]. *Transactions on Machine Learning Research*, 2023.
- [36] Jin Z, Liu J, LYU Z, et al. Can large language models infer causation from correlation? [C]//*The Twelfth International Conference on Learning Representations*. 2024.
- [37] Milianni M, Auriemma S, Bondielli A, et al. Explica: Evaluating explicit causal reasoning in large language models [C]//*Findings of the Association for Computational Linguistics: ACL 2025*. 2025: 17335-17355.

- [38] Yu L, Chen D, Xiong S, et al. Causaleval: Towards better causal reasoning in language models [C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025: 12512-12540.
- [39] Morgan S L, Winship C. Counterfactuals and causal inference: Methods and principles for social research [M]. Cambridge University Press, 2014.
- [40] Sinha K, Sodhani S, Dong J, et al. Clutrr: A diagnostic benchmark for inductive reasoning from text [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 4506-4515.
- [41] Liu N F, Lin K, Hewitt J, et al. Lost in the middle: How language models use long contexts [J]. Transactions of the Association for Computational Linguistics, 2024, 12: 157-173.
- [42] Bai Y, Lv X, Zhang J, et al. Longbench: A bilingual, multitask benchmark for long context understanding [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 3119-3137.
- [43] Wang W, Zhang S, Ren Y, et al. Needle in a multimodal haystack [C]//The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track. 2024.
- [44] Schölkopf B, Locatello F, Bauer S, et al. Toward causal representation learning [J]. Proceedings of the IEEE, 2021, 109(5): 612-634.
- [45] Yao L, Chu Z, Li S, et al. A survey on causal inference [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2021, 15(5): 1-46.
- [46] Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding [C]//International Conference on Learning Representations. 2021.
- [47] Suzgun M, Scales N, Schärli N, et al. Challenging big-bench tasks and whether chain-of-thought can solve them [C]//Findings of the Association for Computational Linguistics: ACL 2023. 2023: 13003-13051.
- [48] Han K, Kuang K, Zhao Z, et al. Causal agent based on large language model [J]. arXiv preprint arXiv:2408.06849, 2024.
- [49] Splawa-Neyman J, Dabrowska D M, Speed T P. On the application of probability theory to agricultural experiments. essay on principles. section 9 [J]. Statistical Science, 1990: 465-472.
- [50] Rubin D B. Estimating causal effects of treatments in randomized and nonrandomized studies. [J]. Journal of Educational Psychology, 1974, 66(5): 688.
- [51] Holland P W. Statistics and causal inference [J]. Journal of the American Statistical Association, 1986, 81(396): 945-960.
- [52] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect [J]. Mathematical Modelling, 1986, 7(9-12): 1393-1512.
- [53] Greenland S, Pearl J, Robins J M. Causal diagrams for epidemiologic research [J]. Epidemiology, 1999, 10(1): 37-48.
- [54] Reynolds L, McDonell K. Prompt programming for large language models: Beyond the few-shot paradigm [C]//Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 2021: 1-7.
- [55] Long S, Schuster T, Piché A. Can large language models build causal graphs? [C]//NeurIPS 2022 Workshop on Causality for Real-World Impact. 2022.

- [56] Kiciman E, Ness R, Sharma A, et al. Causal reasoning and large language models: Opening a new frontier for causality [J]. *Transactions on Machine Learning Research*, 2023.
- [57] Min S, Lyu X, Holtzman A, et al. Rethinking the role of demonstrations: What makes in-context learning work? [C]//*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022: 11048-11064.
- [58] Rubin O, Herzig J, Berant J. Learning to retrieve prompts for in-context learning [C]//*Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022: 2655-2671.
- [59] Liu J, Shen D, Zhang Y, et al. What makes good in-context examples for gpt-3? [C]//*Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. 2022: 100-114.
- [60] Roemmele M, Bejan C A, Gordon A S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. [C]//*AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*. 2011: 90-95.
- [61] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 24824-24837.
- [62] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 22199-22213.
- [63] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models [C]//*The Eleventh International Conference on Learning Representations*. 2023.
- [64] Zhou D, Schärli N, Hou L, et al. Least-to-most prompting enables complex reasoning in large language models [C]//*The Eleventh International Conference on Learning Representations*. 2023.
- [65] Wang L, Xu W, Lan Y, et al. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models [C]//*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023: 2609-2634.
- [66] Besta M, Blach N, Kubicek A, et al. Graph of thoughts: Solving elaborate problems with large language models [C]//*Proceedings of the AAAI Conference on Artificial Intelligence: volume 38*. 2024: 17682-17690.
- [67] Paul D, West R, Bosselut A, et al. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning [C]//*Findings of the Association for Computational Linguistics: EMNLP 2024*. 2024: 15012-15032.
- [68] Chia Y K, Chen G, Tuan L A, et al. Contrastive chain-of-thought prompting [Z]. 2023.
- [69] Zelikman E, Wu Y, Mu J, et al. Star: Bootstrapping reasoning with reasoning [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 15476-15488.
- [70] Zhou P, Pujara J, Ren X, et al. Self-discover: Large language models self-compose reasoning structures [J]. *Advances in Neural Information Processing Systems*, 2024, 37: 126032-126058.
- [71] Shinn N, Cassano F, Gopinath A, et al. Reflexion: Language agents with verbal reinforcement learning [J]. *Advances in Neural Information Processing Systems*, 2023, 36: 8634-8652.
- [72] Zhang C, Zhang L, Wu J, et al. Causal prompting: Debiasing large language model prompting based on front-door adjustment [C]//*Proceedings of the AAAI Conference on Artificial Intelligence: volume 39*. 2025: 25842-25850.

- [73] Chen L, Huang W, Zhang R, et al. A unified causal view of instruction tuning [EB/OL]. 2024. <https://arxiv.org/abs/2402.06220>.
- [74] Li Z, Ding X, Liao K, et al. Causalbert: Injecting causal knowledge into pre-trained models with minimal supervision [EB/OL]. 2021. <https://arxiv.org/abs/2107.09852>.
- [75] Zheng J, Ma Q, Qiu S, et al. Preserving commonsense knowledge from pre-trained language models via causal inference [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 9155-9173.
- [76] Hüyük A, Xu X, Maasch J R M A, et al. Reasoning elicitation in language models via counterfactual feedback [C]//The Thirteenth International Conference on Learning Representations. 2025.
- [77] Vashishtha A, Kumar A, Pandey A, et al. Teaching transformers causal reasoning through axiomatic training [C]//Forty-second International Conference on Machine Learning. 2025.
- [78] Han K, Zhao W, Zhao Z, et al. Cat: Causal attention tuning for injecting fine-grained causal knowledge into large language models [C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 9915-9932.
- [79] Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models [C]//The Eleventh International Conference on Learning Representations. 2022.
- [80] Tang Z, Wang R, Chen W, et al. Towards causalgpt: A multi-agent approach for faithful knowledge reasoning via promoting causal consistency in llms [Z]. 2025.
- [81] Li G, Hammoud H, Itani H, et al. Camel: Communicative agents for” mind” exploration of large language model society [J]. Advances in Neural Information Processing Systems, 2023, 36: 51991-52008.
- [82] Hong S, Zhuge M, Chen J, et al. Metagpt: Meta programming for a multi-agent collaborative framework [C]//The Twelfth International Conference on Learning Representations. 2023.
- [83] Liu Z, Zhang Y, Li P, et al. A dynamic llm-powered agent network for task-oriented agent collaboration [C]//First Conference on Language Modeling. 2024.
- [84] Frohberg J, Binder F. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models [C]//Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022: 2126-2140.
- [85] Qin L, Bosselut A, Holtzman A, et al. Counterfactual story reasoning and generation [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 5043-5053.
- [86] Yu W, Jiang M, Clark P, et al. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions [C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 8276-8288.
- [87] Li J, Yu L, Ettinger A. Counterfactual reasoning: Testing language models’ understanding of hypothetical scenarios [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2023: 804-815.
- [88] Huang Y, Hong R, Zhang H, et al. Clomo: Counterfactual logical modification with large language models [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 11012-11034.
- [89] Chen Y, Singh V K, Ma J, et al. Counterbench: Evaluating and improving counterfactual reasoning in large language models [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 40. 2026: 30350-30358.

- [90] Nie A, Zhang Y, Amdekar A S, et al. Moca: Measuring human-language model alignment on causal and moral judgment tasks [J]. *Advances in Neural Information Processing Systems*, 2023, 36: 78360-78393.
- [91] Mirza P, Tonelli S. Catena: Causal and temporal relation extraction from natural language texts [C]//*Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016: 64-75.
- [92] 邓鹏. 基于语义增强的事件图谱补全和推理研究 [D]. 中国科学院计算技术研究所, 2023.
- [93] 李春尧. 基于模式和约束的知识图谱模型研究与实现 [D]. 中国科学院计算技术研究所, 2018.
- [94] Liu Y, Cao Y, Lin X, et al. Enhancing large language model for knowledge graph completion via structure-aware alignment-tuning [C]//*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025: 20981-20995.
- [95] Yamin K, Gupta S, Ghosal G R, et al. Failure modes of llms for causal reasoning on narratives [J]. *arXiv preprint arXiv:2410.23884*, 2024.
- [96] Romanou A, Montariol S, Paul D, et al. Crab: Assessing the strength of causal relationships between real-world events [C]//*Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023: 15198-15216.
- [97] Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks [C]//*Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.
- [98] Qwen, :, Yang A, et al. Qwen2.5 technical report [Z]. 2025.
- [99] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [C]//*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019: 4171-4186.

致 谢

潺潺流水，终于走过群山一座座。

回望三年的硕士生涯，心中感慨良多。一路走来，有过迷茫与低谷，更多是成长与收获，而每一步前行，都离不开师长的教诲、亲友的陪伴以及同窗的扶持。在论文完成之际，谨以此文，向所有给予我帮助和温暖的人致以最诚挚的感谢。

首先，我要衷心感谢我的导师王石老师。三年前，正值我人生中最困顿的阶段。彼时的我数次考研失利，承受着外界的诸多质疑，也因人生轨迹与同龄人渐行渐远而倍感迷茫，那段时光于我而言格外沉重。正是在这样的时刻，王老师给予了我宝贵的机会，让我得以重新出发，也重新找回前行的方向。

在之后的学习和生活中，王老师始终给予我很多关怀与帮助。王老师为人坦荡豁达，待人谦和宽厚，对学生包容耐心、循循善诱。无论是在科研上的悉心指导，还是在生活中的关心照顾，都给予了我许多支持。老师不仅教会了我如何做学问，更让我从他身上学到了许多受益终身的道理：凡事多思多想，始终以饱满的热情面对生活与工作，脚踏实地、勤勉不辍，坦荡磊落。回顾三年时光，我深知自己仍有许多不足，未能完全达到老师的期许，心中常怀惭愧。能成为王老师的学生，是我人生中莫大的幸运，这份知遇之恩与教诲之情，我将始终铭记于心。

同时，我要感谢实验室的各位老师和同学们。感谢杨竞师兄、孙严顺师兄，在科研和生活中给予我的诸多帮助与指导；感谢同门黄宇威，与你一路互帮互助、并肩前行，这段共同奋斗的同窗情谊弥足珍贵；感谢资康莉老师、裴晓雪老师在学习与生活中给予我的关照与帮助。

此外，我还要感谢我的父母和妹妹。多年来你们给予的理解与支持，始终是我前行路上的底气所在。

感谢我的室友唐智鹏、廖逸之、朱旗、匡万达、王华雷、魏文韬。感谢你们陪伴我度过这段珍贵的研究生时光，你们的陪伴与情谊为这段岁月增添了许多珍贵的记忆。

最后，我要感谢那个曾经身处困顿却从未放弃的自己。三年白驹过隙，计算所见证了我的成长与改变。此去经年，愿自己带着王老师的教诲与这份来之不易的坚韧，在学习和工作中，不畏风雨，砥砺前行。

2026 年 6 月

作者简历及攻读学位期间发表的学术论文与其他相关学术成果

作者简历：

姓名：罗天宇 性别：男 出生日期：1999 年 5 月 2 日 籍贯：贵州省黔东南州

2017 年 9 月——2021 年 7 月，在四川大学高分子科学与工程学院获得学士学位。

2023 年 9 月——至今，在中国科学院计算技术研究所攻读硕士学位。

参加的研究项目及获奖情况：

参与项目情况：

工信部国家重大科技专项《面向生产制造环节的生成决策工业大模型》(CEIEC-2024-ZM02-0055) 子任务《生产制造工业大模型中的长程多步决策关键技术研究与应用》

获奖情况：

2024-2025 学年 中国科学院大学三好学生

2024-2025 学年 中国科学院大学优秀学生干部

2023 年中国科学院大学学业奖学金

2024 年中国科学院大学学业奖学金

2025 年中国科学院大学学业奖学金

