北京交通大学

硕士学位论文

融合小型专家模型的大语言模型推理增强研究

Inference Enhancement for Large Language Models via Small Expert Models

作者：张京

导师：桑基韬

北京交通大学

2025 年 6 月

学校代码: 10004 密级: 公开

北京交通大学

硕士学位论文

融合小型专家模型的大语言模型推理增强研究

Inference Enhancement for Large Language Models via Small Expert Models

作者姓名: 张京 学　　号: 22120464

导师姓名: 桑基韬 职　　称: 教授

学位类别: 工学 学位级别: 硕士

学科专业: 计算机科学与技术 研究方向: 自然语言处理

北京交通大学

2025 年 6 月

摘要

近年来，大语言模型（Large Language Models, LLMs）在通用推理与自然语言理解方面表现出色，应用场景不断拓展。然而，由于专业知识覆盖有限、更新滞后等问题，LLMs 在垂直领域上的应用仍存在不足。相比之下，小型专家模型（Small Expert Models, SEMs）在特定领域具备更强的适应性与专业性，可有效弥补 LLMs 在领域知识方面的短板。但现有协同方法面临两大挑战：一是在有真实标签场景下，将存在偏差的 SEMs 概率作为置信度容易误导 LLMs，同时在 SEMs 预测错误时缺乏相应的优化策略；二是在无标签场景中，依赖 SEMs 偏好生成的知识缺乏有效的验证机制，容易引入错误信息。

为应对上述挑战，本文在有真实标签与无真实标签两类场景下，系统探讨了大小模型协同推理机制，并在自动报告生成场景中进行应用，主要研究内容包括：

(1) 基于概率校准与错误驱动的有标签协同推理算法。针对置信度偏差问题，本文首先采用概率校准模型对 SEMs 输出的概率进行调整以纠正偏差。同时，引入错误驱动机制，提取错误样本中的推理规律，帮助 LLMs 在 SEMs 发生错误时，也能获得针对性的修正策略。实验证明，该方法在新闻分类、生物医学问答等四个数据集上显著提升了准确率、F1 值和鲁棒性。

(2) 基于多源知识验证的无标签协同推理算法。为缓解错误知识扩散风险，本文构建全局与局部双粒度的知识验证体系。在全局层面，利用聚类提炼稳健的领域共性知识；在局部层面，构建多假设候选集，通过 LLMs 决策获得更具说服力的局部知识，提升知识质量。实验结果表明，该方法在多任务、多设置条件下均有效增强了 LLMs 的推理能力，并具备良好的鲁棒性。

(3) 面向自动报告生成的大小模型协同框架。本文提出 AutoRG 框架，实现从大纲规划、内容生成到成稿的全流程自动化报告生成。采用先召回后排序的调用方式缓解多 SEMs 调用不稳定的问题；通过信息修正与报告迭代流程，缓解 SEMs 误差和 LLMs 输出波动，提升内容准确性与逻辑一致性。实验验证，AutoRG 显著提升报告的整体质量与稳定性，具有良好的实用性与推广前景。

关键词：大语言模型；协同推理；小型专家模型；大模型推理增强；智能体

ABSTRACT

In recent years, large language models (LLMs) have demonstrated impressive capabilities in general inference and natural language understanding, leading to a growing range of application scenarios. However, due to limited coverage of domain-specific knowledge and delayed updates, their performance in vertical domains remains suboptimal. In contrast, small expert models (SEMs), with their superior adaptability and specialization in specific fields, serve as effective complements to LLMs in compensating for domain knowledge gaps. Nonetheless, existing collaboration methods face two major challenges: (1) in scenarios with ground-truth labels, biased confidence scores from SEMs may mislead LLMs, and there is no corresponding optimization strategy when SEMs are wrong; (2) in scenarios without ground-truth labels, knowledge generated based on SEM preferences often lacks effective verification mechanisms, increasing the risk of erroneous information propagation.

To address these challenges, this study systematically investigates collaborative inference mechanisms between LLMs and SEMs under both labeled and unlabeled scenarios and applies them to the task of automated report generation. The main contributions are as follows:

(1) Supervised Collaborative Inference Algorithm Based on Probability Calibration and Error-Driven Mechanisms. To correct confidence bias, this paper first employs a probability calibration model to adjust the confidence scores of SEM outputs. Additionally, an error-driven mechanism is introduced to extract reasoning patterns from error samples, enabling LLMs to obtain targeted correction strategies even when SEMs are worong. Experimental results on four benchmark datasets—including news classification and biomedical question answering—demonstrate significant improvements in accuracy, F1 score, and robustness.

(2) Unsupervised Collaborative Inference Algorithm Based on Dual-Source Knowledge Verification. To reduce the risk of erroneous knowledge propagation, this paper proposes a dual-source knowledge verification framework. At the global level, clustering is used to extract robust domain-level common knowledge; at the local level, a multi-hypothesis candidate set is constructed, and more persuasive local knowledge is obtained through LLM-based decision-making, thereby improving overall knowledge quality. Experimental results demonstrate that this method effectively enhances the reasoning capabilities of LLMs across multiple tasks and settings and has good robustness.

(3) LLM-SEM Collaborative Framework for Automated Report Generation. This paper proposes the AutoRG framework to enable a fully automated report generation process, from outline planning and content generation to final drafting. A recall-then-rank strategy is adopted to alleviate instability when invoking multiple SEMs. Furthermore, an information correction and report iteration mechanism is introduced to mitigate SEM errors and LLM output fluctuations, improving both content accuracy and logical consistency. Experiments confirm that AutoRG significantly improves the overall quality and stability of generated reports, demonstrating strong practicality and potential for deployment.

KEYWORDS: Large Language Models; Collaborative Inference; Small Expert Models; Inference Enhancement for Large Language Models ; Agent

目　录

1 引言

1.1 研究背景与意义

近年来，大语言模型（Large Language Models, LLMs）凭借超大规模的参数量和海量训练数据，在通用自然语言处理任务中表现出接近人类的语义理解和生成能力。以 GPT-4 和 PaLM-2 为代表的 LLMs 不仅推动了文本生成、知识推理等核心技术的发展，还被视为实现通用人工智能的重要路径。

与此同时，提升 LLMs 在特定领域的推理能力对企业具有深远意义。在法律[1,2]、医疗[3–5]、金融[6,7] 等垂直领域中，具备较强领域推理能力的 LLMs 能够帮助企业更精准地解读领域数据和信息，从而优化决策流程、降低误判风险，为企业赢得竞争优势并带来显著的运营效益。然而，在实际应用中，由于 LLMs 在领域知识的覆盖上存在不足，且动态更新能力有限，导致实际部署效果往往无法达到预期的理论水平。

为了解决 LLMs 在垂直领域的适配问题，现有研究主要沿两大技术路线展开探索：一类是参数微调（Fine-tuning）方法，利用领域数据对 LLMs 进行有针对性的训练，从而提升其在特定任务中的表现；另一类则是在保持 LLMs 参数不变的前提下，通过推理阶段的增强（Inference Enhancement）手段，引导 LLMs 更充分地利用已有知识或结合外部信息进行领域推理。在企业级场景中，如何以更低的成本、更高的效率完成大模型高质量、稳定输出是实际部署时面临的难题。与前者相比，推理增强方法在灵活性、适配成本和可控性方面具有天然优势，特别适用于难以获得大量标注数据或对模型稳定性要求较高的实际应用场景。因此，本文聚焦于 LLMs 的推理增强，旨在探索如何在推理阶段以更高效、低成本的方式引入领域知识，从而提升 LLMs 在领域任务中的推理能力。

LLMs 在推理阶段的增强方式根据所引入知识的来源分为两类：基于内部知识的提示引导增强与基于外部知识的知识注入增强。提示引导增强主要依赖提示学习技术，如上下文学习（In-Context Learning, ICL）[8] 和思维链（Chain-of-Thought, CoT）[9]，通过精心设计的提示词，引出 LLMs 在预训练时就已内化的知识，实现对任务的推理。然而，由于 LLMs 在预训练时固有的知识存储局限性以及知识更新滞后问题，这种方法在面对快速变化的领域知识时存在一定的局限。为弥补这一不足，研究者开始探索知识注入增强方式。一方面，检索增强方法通过引入与当前任务相关的外部文档或知识片段，为 LLMs 提供即时的上下文补充[10,11]，提升其在专业领域的表现；另一方面，基于智能体（Agent）框架的方法进一步扩展

了 LLMs 的能力边界。通过集成计算器[12]、API 接口[13,14] 等外部工具，LLMs 可专注于任务分解与推理过程，而将具体计算或领域任务的执行交由专业组件完成，从而实现更强的模块化协同。虽然知识注入增强的方式有效地实现了 LLMs 与知识的解耦，但仍面临知识匹配不准确以及相关知识库和工具收集难度较大的问题。

在此背景下，小型专家模型（Small Expert Models, SEMs）与 LLMs 的协同机制逐渐成为领域适配研究的新趋势。当前的 LLMs 与 SEMs 的核心差异被归结为“LLMs 更通，SEMs 更精”。即，LLMs 在多任务、多领域的通用推理上具备显著优势；而 SEMs 则聚焦于特定任务或领域，训练目标更集中，推理结果更稳定可靠。因此，融合 SEMs 以增强 LLMs 推理能力的框架旨在通过构建一种互为补充、协同增强的机制，实现以小补大、以专助通的目标，在确保任务性能的前提下维持通用能力不退化的同时，提升系统整体的应用价值和可靠性。在现有的大小模型协同框架中，一种较为常见的方式是构建从 SEMs 到 LLMs 的信息传递通道，实现与 LLMs 的推理框架的无缝融合。

然而，目前的大小模型协同方法依然存在明显缺陷。一方面，在有真实标签的场景下，主流方法倾向于将 SEMs 的预测标签和真实标签共同嵌入提示模板，通过混合的标签信号引导 LLMs 做出正确推理。但由于 SEMs 置信度存在偏差，LLMs 往往难以准确判断其输出的可靠性，易导致误判。此外，当前方法也缺乏在 SEMs 预测错误后的优化机制，限制了 LLMs 的纠错能力。另一方面，在无真实标签的场景中，部分研究试图利用 SEMs 的偏好信息构建知识库，并利用检索增强来扩展 LLMs 的领域知识。然而，由于知识生成过程缺少稳定而可靠的验证机制，导致单一依赖 SEMs 构建的知识库容易引入错误信息，并将这些错误知识传播给 LLMs，从而产生误导效应。这两类问题严重制约了现有协同框架的推理能力，因此需要进一步完善 SEMs 与 LLMs 间的信息传递过程。

1.2 主要研究内容

本研究聚焦融合 SEMs 的 LLMs 推理增强，研究框架如图1-1所示。

一方面，本研究在理论层面优化 SEMs 与 LLMs 在信息传递过程中的协同推理机制，旨在解决两类关键问题：其一是在有真实标签场景中解决 SEMs 置信度偏差及预测错误时缺乏优化策略的问题；其二是在无真实标签场景中构建可靠的知识生成机制，解决知识生成不准确的问题。另一方面，本研究还在实践层面对基于 Agent 的工具增强框架展开探索，以自动报告生成为场景，验证多 SEMs 与 LLMs 协同下的应用价值。具体研究内容包括以下三个维度：

(1) 基于概率校准与错误驱动的有标签协同推理算法：针对 SEMs 输出置信度偏差及缺乏 SEMs 预测错误后的优化策略的问题，本研究提出一种融合概率校准

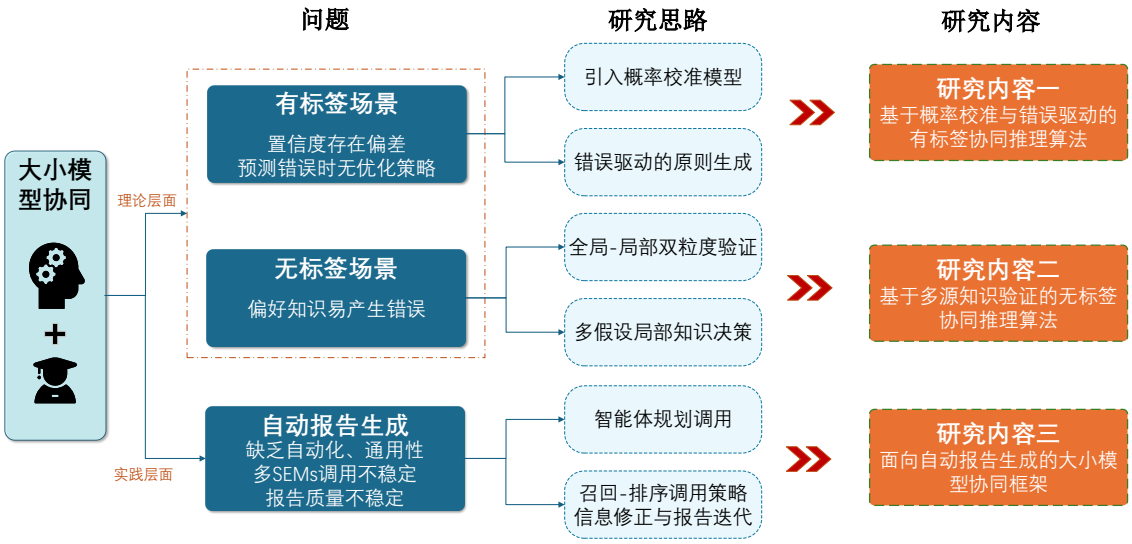


图 1-1 融合 SEMs 的 LLMs 推理增强框架

Fig. 1-1 Inference Enhancement for Large Language Models via Small Expert Models

与错误驱动的联合推理方法。首先，通过采用 Beta 概率校准技术对 SEMs 输出的概率进行置信度量化，纠正 LLMs 在预测过程中出现的错误偏好；其次，构建专门的错误分析模块，利用 SEMs 的历史错误样本生成错误规避原则，并将该原则注入 LLMs 的推理提示中，从而在 SEMs 预测错误的场景下引导 LLMs 生成更为准确和合理的回复。

(2) 基于多源知识验证的无标签协同推理算法：为降低错误知识被传递的风险，本研究设计了全局与局部双粒度的知识验证体系。在全局层面，通过数据聚类方法对预测数据进行分组，随后对其采样以生成全局知识，获得更鲁棒的粗粒度知识；在局部层面，提出多假设的局部知识决策算法，对 SEMs 输出的多个候选偏好知识进行投票，选取更具说服力的见解作为局部知识，从而确保局部知识的准确性。

(3) 面向自动报告生成的大小模型协同框架：最后，本研究以自动技术报告生成为实际应用场景，验证多 SEMs 与 LLMs 协同的有效性。该框架充分利用 LLMs 在工具理解与自主规划方面的优势，实现技术报告的自动化生成；同时，采用先召回后排序的方式优化多 SEMs 调用机制，并引入信息修正与报告迭代机制，有效降低生成的报告对 SEMs 输出和 LLMs 随机性的依赖，从而提升报告生成的质量和稳定性，确保生成的内容兼具逻辑的严谨性与领域的精准性。

1.3 论文组织结构

本文围绕融合 SEMs 的 LLMs 推理增强展开系统性研究，全文共分为六章，各章节内容安排如下：

第一章为引言，主要介绍大模型领域推理的研究背景与意义，回顾大小模型协同相关研究现状，明确本文的研究内容、目标与创新点，并对全文的整体架构与安排做出概述，为后续章节奠定理论基础和研究动机。

第二章系统梳理国内外在 LLMs 推理增强方面的研究工作。章节首先从提示引导增强与知识注入增强两个视角探讨了 LLMs 在推理增强领域中的进展；随后对 SEMs 在协同增强框架中所处的不同角色与阶段进行归纳，总结其在辅助 LLMs 完成各类任务中的应用；最后，重点分析了知识交互框架下的各种方法与实践，为后续研究提供理论支持与启示。

第三章详细描述了基于概率校准与错误驱动的有标签协同推理算法。章节首先分析了问题背景，并阐述了研究该方向的动机；接着从方法论角度出发，详细介绍了 Beta 概率校准模块与错误分析等模块的设计思路与实现机制；随后描述了实验设计、参数设置及对比基线；最后，通过一系列实验结果，对不同配置下的性能表现进行系统性分析与深入对比，验证所提方法的有效性。

第四章着重介绍了基于多源知识验证的无标签协同推理算法。章节开篇进行问题剖析并明确研究动机，随后分别介绍了从粗粒度到细粒度的知识生成方法及其应用机制，详细描述了全局-局部双粒度知识验证体系的构建过程；接下来说明了实验配置与对比方法；最后，围绕各实验结果展开全面的对照分析，探讨了该方法在提升 LLMs 推理任务方面的优势。

第五章主要介绍了面向自动报告生成的大小模型协同框架。首先回顾自动报告生成领域的相关研究，并指出当前存在的问题与不足；接着详细描述了所提出协同框架的组成结构与执行流程，阐明 LLMs 与 SEMs 之间协同工作的方法逻辑；最后，通过主观与客观两个维度设计实验进行评估与分析，验证协同推理框架在自动报告生成任务中所展现的高质量和鲁棒性。

第六章总结全文工作，梳理主要研究成果与创新点，并从实验结果与理论意义上对全文进行综合评价。同时，针对当前方法存在的局限性，探讨未来可能的改进方向与进一步的研究领域，为后续工作提供展望与思路。

2 国内外研究现状

2.1 大模型推理增强

近年来，大语言模型（LLMs）在复杂推理任务中展现出突破性进展，逐步被视为通用任务求解的重要基础。但实际上，将其从通用语言理解扩展至具备专业推理能力的领域化应用，仍面临诸多挑战。由此，如何更好的引入领域知识以增强 LLMs 推理能力逐渐成为研究关注的重点。根据知识的来源不同，可把增强方式分为基于内部知识的提示引导增强与基于外部知识的知识注入增强两类[15]。

2.1.1 提示引导增强

LLMs 凭借其庞大的参数量实现丰富的内部信息存储，从而生成高质量的响应，但这些响应并不总能确保包含准确的知识。因此，如何促使 LLMs 输出正确的隐式知识成为亟待解决的问题，这就需要对输入的提示进行额外的处理和优化。

提示工程是一种常见的 LLMs 推理策略，其核心在于通过精心设计、改进和优化输入指令或 ‘‘提示’’，使模型产生更加准确、相关且可控的输出。传统的全模型微调通常需要对全部模型参数进行更新，随着模型规模不断扩大，其效率瓶颈愈发明显。与传统的全模型微调方法相比，提示工程仅通过调整任务的呈现方式，即可激发 LLMs 中已编码的潜在知识，避免了大规模参数更新所带来的高昂计算成本，为大模型的高效部署提供了可行路径。

早期研究主要关注上下文学习。零样本提示（Zero-Shot Prompting）通过构建与任务紧密相关的自然语言提示，使 LLMs 在没有额外数据的情况下完成推理工作[16]。这种方法能够充分利用 LLMs 的先验知识，实现对未知任务的泛化，但对提示设计的依赖性较强。Monica[17] 在临床任务中采用零样本提示，测试结果表明其性能超越了多种传统方法。

相比之下，少样本提示（Few-Shot Prompting）进一步拓展了零样本提示的应用，通过在输入中嵌入高质量的示例，增强了模型根据上下文信息进行推理的能力。Brown 等人[8] 的实验证明，即便是提供少量示例，也能显著提高 GPT-3 在复杂任务上的表现，尤其在数学推理和常识问答领域中展现出优异的能力，这一策略也被广泛应用于多个领域的任务中[18]。例如，Ferber 等人[19] 利用上下文学习成功赋予多模态 LLMs 进行癌症病理学图像分类的能力；Mo 等人[20] 则提出了基于对比示例的 C-ICL 方法，显著增强了 LLMs 对信息抽取任务的理解和处理效果。

然而，需要指出的是，上下文学习不总是优于零样本学习，其效果往往依赖于

具体任务的特点[21]。此外，示例检索算法与示例数量[22]、排序方式[23]以及标签的准确性[21]等因素也会对 LLMs 的推理性能产生影响。这些都导致上下文学习在实际应用中存在不稳定性问题。因此，设计出既鲁棒又具有良好解释性的提示策略，仍然是推动 LLMs 提示引导增强的重要研究方向。

为提升推理的可控性与可解释性，Wei 等人[9]提出了 CoT 方法。该方法通过显式生成中间推理步骤，引导 LLMs 按逻辑顺序逐步推理，从而实现了对复杂推理过程的分解和控制。在数学和常识推理基准测试中，该方法使 LLMs 的准确率达到了 90.2%，被广泛认为是复杂推理的重要里程碑。与此同时，Liu 等人[24]探索了将化学生物学专家的思维过程融入提示的方法，在有机小分子、酶和晶体材料等领域的数据集上，均显著优于传统的通用提示策略，取得了更优的性能表现。此外，Tree of Thoughts[25]方法在 CoT 的基础上进行了扩展，将整个推理过程拆解为树状结构，允许模型并行探索不同的推理路径，从而在 24 点游戏和创意写作等任务中展现出显著优势。类似地，还有基于图的 GoT[26]和基于渐进分析的 ThoT[27]等方法，这些工作都为提升 LLMs 推理过程的可控性和解释性提供了多样化的技术路径。

2.1.2 知识注入增强

尽管 LLMs 内部蕴含大量知识，但其更新周期长，且重新预训练所需成本高昂，限制了知识的实时性与实用性。此外，上下文学习和基于思维链的推理方法对提示设计依赖较强，性能稳定性仍存在不足。Kandpal 等人[28]指出，LLMs 在处理长尾知识方面存在显著局限，导致其在细分领域的推理能力难以满足实际需求。因此，引入外部知识以增强 LLMs 的推理能力成为研究热点，主要包括基于知识的增强与基于工具的增强两种路径。其中，前者通常通过检索增强实现，后者则以智能体为核心进行建模。

(1) 基于检索的知识增强

检索增强（Retrieval-Augmented Generation, RAG）通过集成外部知识库，有效弥补了 LLMs 在静态知识和长尾知识上的不足，显著提升了其在事实性和时效性方面的推理能力。该类方法通常包括三个关键步骤：首先，LLMs 根据输入问题生成查询语句；其次，通过搜索引擎或领域知识库检索相关文献或文档；最后，对检索结果进行融合与分析，以辅助生成最终答案。这种机制不仅扩展了 LLMs 的知识覆盖范围，还显著提高了其应对复杂、开放性问题的能力。

Lewis 等人[29]提出的 RAG 模型首次实现了非参数化检索与生成的端到端集成，在 TriviaQA 数据集上取得了 68% 的准确率。阿里团队的 Li 等人[30]将问题划分为可直接回答与长尾问题两类，并针对长尾问题引入定向文档检索增强策略，

从而显著提升了推理性能。Yu 等人[31] 开发的 KAR 系统结合知识图谱以优化语义检索，在学术论文和医学问答等场景中表现出领先效果。随着 LLMs 参数规模的持续扩大，检索增强也逐渐向轻量化方向演进。Asai 等人[32] 提出的 Self-RAG 通过引入自我反思机制，动态触发信息检索，有效减少冗余调用，在长文本生成任务中，对事实性与引用准确性方面带来了显著收益。

尽管如此，该方向仍面临若干挑战。一方面，从模型层面来看，检索算法精度提升[32,33]、多源知识冲突协调[34] 以及模型内外部知识间的不一致性[35,36] 等问题仍亟待深入研究。另一方面，随着知识库规模不断扩大，计算成对知识间相似性的成本显著上升，如何构建高效、可扩展的检索增强体系已成为当前研究的关键课题。

(2) 基于智能体的工具增强

智能体是人工智能领域中的一个核心概念，指能够感知自身所处环境、自我决策并采取行动的智能系统[37]。随着 LLMs 推理能力的显著提升，基于 LLMs 的智能体架构展现出巨大的应用潜力。在该框架下，智能体通常以预训练的 LLMs 作为核心推理引擎，辅以模块化组件（如计划规划、记忆管理、工具使用和反馈机制），在没有人工干预的情况下完成高层次、目标驱动的任务。得益于 LLMs 强大的零样本指令跟随能力，智能体能够通过理解工具的自然语言描述，掌握其使用方法，从而完成复杂任务。例如，Toolformer[13] 展示了 LLMs 如何通过调用外部搜索工具来获取实时信息，突破了模型参数中静态知识的限制。

工具是 LLMs 获得知识的关键，常见可集成的工具包括搜索引擎[13]、计算器[12]、代码解释器[14] 及其他辅助模型等。通过实现对工具的高效调用，LLMs 相当于外接了代理，可以可插拔地实现精确的算术、实时信息检索等功能，这些功能超越了 LLMs 参数中本身编码的静态知识。ChatLaw[38] 是一个典型案例，其为法律领域定制的 LLMs 智能体系统，结合数据库与关键词搜索策略，有效缓解了法律问答中常见的幻觉问题，展现出强大的领域问答能力。

此外，智能体的多步推理能力进一步增强了其在复杂任务中的规划与执行能力[37]。规划使智能体能够将复杂任务分解为子任务[39]，并制定合理的行动计划以更加高效地完成长期目标。在智能体协同方向，一些研究进一步探索多个智能体之间的协作机制。MetaGPT[40] 合理利用多个智能体之间的协作，使其在各种应用场景中更加有效和可靠。

2.2 SEMs 辅助的 LLMs 增强

LLMs 凭借其强大的内部知识表达与复杂推理能力，在通用人工智能领域展现出巨大潜力。然而，在实际推理增强框架的部署中，为 LLMs 系统性地收集、筛

选和配置领域知识的过程面临诸多挑战。一方面，LLMs 的预训练语料往往偏向通用领域，难以覆盖专业领域的长尾知识；另一方面，基于检索增强等机制在很大程度上依赖高质量且结构化的知识文档，这对知识工程、数据清洗等环节提出了较高要求[29]。当知识文档分布较为离散时，LLMs 对相关信息的记忆能力也会显著受限[41]，从而使得检索增强在实际应用中难以保障知识的完备性与时效性。

与 LLMs 所需的广泛通用知识和任务适应能力不同，SEMs 通常专注于特定领域中明确定义的任务与领域知识，具备结构轻量、成本可控、可快速部署等优势。因此，融合通用 LLMs 与 SEMs，成为提升模型整体效能与灵活性的关键路径。现有研究主要从三个方面展开探索：知识迁移以提升 SEMs 能力、合作推理以实现动态协作，以及增强 LLMs 的领域适应性。

知识迁移是提升 SEMs 领域表现的重要手段，特别适用于 LLMs 在通用任务中展现出优势但执行成本较高的情境。该方向主要包括三类典型方法：首先，渐进式推理链方法[42,43] 通过将 LLMs 的复杂推理过程拆解为多个中间步骤，引导 SEMs 逐步学习推理结构与逻辑路径；其次，上下文蒸馏策略[44,45] 基于少样本提示能力，将 LLMs 在上下文学习中的优势迁移至轻量化模型，实现 SEMs 在低资源场景下的有效泛化；最后，概率分布模仿方法[46,47] 通过对齐输出概率分布，使 SEMs 更准确地捕捉 LLMs 的生成偏好，从而提升整体性能。

合作推理研究关注如何在推理过程中实现大小模型的动态协作，形成优势互补的推理模式。典型的工作如 SwiftSage 框架[48]，将 LLMs（Sage）负责深度推理与规划，SEMs（Swift）负责快速响应与执行，在复杂交互场景中实现推理精度与响应效率的平衡。Tan 等人[49] 则提出使用轻量级代理模型 SlimPLM 动态评估 LLMs 的知识盲区，仅在必要时触发检索增强，从而减少不必要的计算成本，相比基于 LLMs 自省策略展现出更高的效率与控制性。

在领域适应性增强方面，SEMs 作为可插拔的专家单元，被灵活嵌入至 LLMs 的推理流程中，从数据预处理、知识交互到反馈优化等关键阶段对 LLMs 进行辅助和增强。如图2-1所示。相较于微调方法，直接在推理阶段嵌入 SEMs 具有部署灵活、训练开销小等优势。因此，接下来的部分将围绕直接推理场景下，如何通过 SEMs 提升 LLMs 的领域推理能力进行深入探讨。

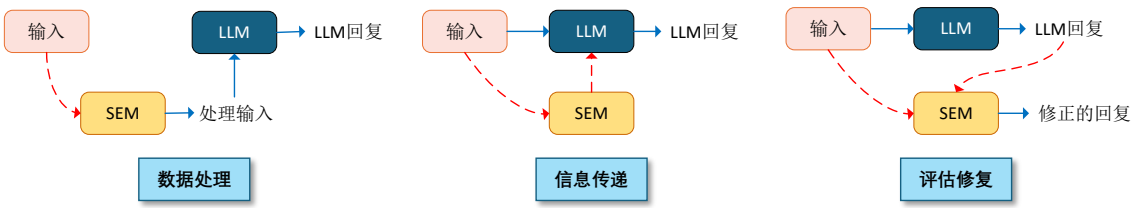


图 2-1 SEMs 增强 LLMs 领域适应性的三种方式

Fig. 2-1 Three Approaches for Enhancing LLMs' Domain Adaptability via SEMs

2.2.1 数据处理

在 LLMs 推理流程的前期阶段，数据质量对模型性能具有显著影响。为此，部分研究尝试在数据预处理环节引入 SEMs，通过其领域特化能力筛选出高信息密度的样本，从而构建更适配 LLMs 推理的数据分布空间。

Mekala 等人[50] 验证了 SEMs 在数据选择中的判别能力，提出基于预测概率分布差异的评价指标，用以衡量样本对 LLMs 性能的关键程度，实验证明，SEMs 能够有效识别出对提升 LLMs 表现具有关键作用的核心训练样本。Li 等人提出的 Superfiltering 框架[51] 构建了由 ‘‘弱’’ 到 ‘‘强’’ 的多级数据过滤流程：该方法首先利用 SEMs 计算数据的困惑度与指令跟随难度得分，再结合模型间排序一致性，保留 10%-30% 的高质量样本，显著提升数据密度与训练效率，使 LLMs 在保持性能不损的前提下微调速度提高 5 至 10 倍。此外，LESS 方法[52] 从任务特定性的角度优化样本选择。该方法引入梯度相似性指标，度量训练样本对目标验证集的性能贡献，并据此选出最具代表性的子集。在数学推理与代码生成等任务中，LESS 使用仅 20% 的数据即可恢复超过 95% 的全量训练性能，显著降低了数据使用成本。

2.2.2 信息传递

在知识交互阶段，可构建从 SEMs 向 LLMs 的信息传递机制。根据是否依赖领域标签数据，现有研究主要分为两类：有标签场景下的知识融合与无标签场景下的检索增强。

在有标签场景中，研究者通常借助提示工程（Prompt Engineering）促进 LLMs 与 SLMs 之间的协同，以提升任务性能或缓解分布外泛化能力不足的问题。Super-ICL[53] 通过将 SEMs 的预测标签及其置信度拼接至输入的上下文提示中，引导黑盒 LLMs 优化监督分类任务的输出。然而，其性能受限于 SEMs 的预测质量，且数据改写率偏低。相比前者对知识可信度的侧重，SuperContext[54] 更关注于对 SEMs 错误的修正，利用 SLMs 的输出作为 ‘‘监督知识’’ 嵌入提示，显式干预 LLMs 的生成逻辑。该方法允许 LLMs 评估监督信息的可靠性，并动态决定是否采纳。但它们都依然严重依赖置信度的准确性，且在 SEMs 预测错误的情境下缺乏后续优化策略。

尽管在部分数据集和特定 SEMs 的支持下，借助有真实标签的数据可以实现超越 SEMs 的性能，但对于大多数任务来说，获取标签信息仍然非常困难。因此，在无真实标签场景下，一些研究者尝试通过检索增强的方法进行偏好对齐，从而构建出更轻量的知识迁移机制，以低成本的方式将领域知识注入 LLMs。BLADE[11] 聚焦法律、医疗等垂直领域，提出三阶段迁移机制：领域预训练（DP）、知识指导

微调（KIT）与贝叶斯提示优化（BPO），使 SEMs 具备领域表达能力并在推理时调整与 LLMs 的输出对齐。实验证明其在法律问答任务中显著提升 ChatLaw-13B 的表现（提升达 67.8%），验证了 SEMs 辅助适配的有效性。与 BLADE 依赖领域语料的方式不同，PANDA[10] 跳过参数更新，提出从 SEMs 的概率分布中提取偏好对，引导 LLMs 生成解释性见解并构建知识池，并通过检索增强将偏好迁移至 LLMs。该方法不仅具备较强解释性，还在 ScienceWorld 任务中实现了对 SEMs 性能的超越，揭示了 SEMs 在引导 LLMs 泛化能力释放方面的潜力。然而，其性能高度依赖于 SEMs 偏好的准确性与一致性。

此外，还有一类工作关注弱到强泛化机制（Weak-to-Strong Generalization）[55]。其核心思想在于通过向较弱模型对齐来激发强模型原有的潜在能力，使得即便弱模型仅能提供不完整或存在错误的训练标签，强模型依然能够在泛化上取得超越。OpenAI 超级对齐团队[55] 利用 GPT 系列模型，在 22 个 NLP 任务、棋类任务和奖励模型分类任务上系统性验证了 ‘‘弱标签可激发强模型潜在能力’’ 的设想。Aligner[56] 通过学习未对齐答案与对齐答案之间的修正残差，进一步提出了一种替代传统强化学习人类反馈（RLHF）的新方法，其理论思想为 SuperICL 和 Super-Context 工作提供了理论基础。与此同时，Liu 等人[57] 引入了多样化的教师模型集群，通过协同监督强模型，以缓解单一教师模型在任务微调后性能受限的问题。该方法借鉴了混合专家框架，并在 OpenAI 的 Weak-to-Strong 基准测试以及多个领域的数据集上验证了其有效性。虽然弱到强泛化所使用的 SEMs 更趋向于通用型而非专家型，强调语言模型规模的大小，但其多教师协作与修正学习的思想同样适用于构建 SEMs 辅助的 LLMs 推理增强框架。

2.2.3 评估修复

SEMs 与 LLMs 之间的协作不仅限于前端的推理过程，还包括对输出结果的后处理和校准，以进一步提高生成内容的准确性和可信度。

一种常见的策略是利用级联框架：在此框架下，SEMs 首先对输入数据进行领域适配，接着 LLMs 负责生成主体推理结果，最后 SEMs 再次对 LLMs 的输出进行校准，从而实现输出内容的进一步优化。CombLM[58] 提出的该级联增强方法在医疗文本生成任务中使 BLEU 分数提升了 18.2%。在更复杂的校准框架中，SEMs 会结合输入问题和 LLMs 生成的答案，通过最小化估计的校准误差与预测置信度之间的差距，动态调整其输出结果。Zhao 等人[59] 利用一个轻量级的 BERT 模型对每个 LLMs 响应生成风险评分，从而系统地校准模型响应。该方法在生物医学和一般领域的标准关系抽取任务中均实现了超过当前最先进结果的性能。

此外，还有一些工作关注利用 SEMs 进行幻觉检测，通过分析 LLMs 内部的状

态信息，如激活状态和注意力分布，来判断生成内容中可能存在的幻觉问题。Xu 等人[60] 开发了一种轻量级幻觉检测器，能够实时识别 LLMs 输出中的异常模式，并用于辅助纠正神经机器翻译中的幻觉现象。该方法在无需依赖额外外部资源的情况下，实现了高效的检测与修正，进一步提升了输出结果的稳定性与可靠性。

2.3 本章小结

本章系统梳理了国内外在 LLMs 推理增强领域的研究进展，主要聚焦于提示引导增强及知识注入增强。提示引导增强依赖优化的输入提示，以激发 LLMs 内部潜在知识，涵盖零样本提示、少样本提示、思维链（CoT）及其扩展策略等方法。知识注入增强则侧重于引入外部知识，通过检索增强与智能体工具调用的方式，弥补 LLMs 在知识广度、时效性及专业性方面的不足。

为进一步提升 LLMs 的推理效率与领域适应能力，融合 SEMs 与 LLMs 的研究逐渐兴起。相关工作主要围绕知识迁移、合作推理与领域适应性展开。其中，领域适应性增强聚焦于 SEMs 在推理流程中作为专家模块参与数据选择、知识注入及输出校准的过程，提升 LLMs 在专业领域的稳定性与可信度。此外，近年来研究还扩展至弱到强泛化，通过多教师协同与输出后处理进一步挖掘模型潜能，构建出更具弹性与可信度的推理增强框架。综上，SEMs 在辅助 LLMs 方面展现出重要价值，其融合机制已成为推动 LLMs 领域化的重要研究路径。

3　基于概率校准与错误驱动的有标签协同推理算法

3.1　引言

在有真实标签的场景下，如何有效融合 LLMs 与 SEMs 的协同推理能力，已成为提升垂直领域任务性能的研究方向之一。现有研究[53,54] 提出双标签通道框架，通过将 SEMs 预测的标签（y'）与人工标注的真实标签（y）共同嵌入提示模板，在 LLMs 的上下文学习过程中建立预测结果与真实目标之间的映射。该技术路径与 LLMs 对齐研究[61] 中提出的弱监督信号转化机制在本质上具有一致性，即相较于直接学习从查询到答案的复杂映射，利用上游模型输出构建修正任务更具有操作性和可解释性。

然而，现有方法存在以下两个问题。首先，SEMs 输出的原始概率作为置信度会存在系统性分布偏差 。如图3-5-a) 所示，当直接采用未经校准的预测概率作为可靠性指标时，SEMs 呈现出明显的概率误差。在置信度在 0.9-0.95 的高概率区间，实际预测准确率仅不到 45%（过度自信现象）；而在 0.8-0.85 的中概率区间，实际正确样本占比却超过 50%（预测保守现象）。这种偏差本质上是一种域分布偏移问题。在测试域中，数据分布可能与训练集不同，导致模型过度拟合训练数据，无法泛化真实的不确定性[62]。近期研究[63] 表明，通过增强弱监督信号的可靠性能够显著改善弱到强的对齐效果，这为本研究设计概率校准机制以获得更准确的可靠性衡量标准提供了理论依据。通过引入概率校准机制，有望优化预测结果的置信度，进而改善 LLMs 在垂直领域中的推理效果。

其次，当 SEMs 发生错误的时候，缺乏相应的优化策略。已有的提示信息仅包含 SEMs 的预测标签与概率，无法直接为 LLMs 提供错误示例情况下的正确推理原则，导致 LLMs 仅能依据自身能力进行简单的拒绝采信判断，推理性能存在不足。值得注意的是，从错误中学习的理念在机器学习领域已有长时间探索，例如早期 Minsky 的感知机理论[64] 就强调错误样本对决策边界修正的重要性。近期研究[65,66] 进一步验证了错误原则在上下文学习框架中的有效性，这启发本研究构建动态错误学习机制。即，对于 SEMs 更有可能预测正确的高置信度预测案例，LLMs 可直接采纳 SEMs 的结果；而对于容易出错的低置信度案例，则通过分析历史错误模式，提炼出指导性推理原则，从而在整体上提升系统的推理性能。

针对上述问题，本研究提出基于概率校准与错误驱动的有标签协同推理算法（Calibrated Confidence and Diagnostic Contextualization, CCDC），主要贡献如下：

(1) 引入基于概率校准的置信度评估模型。设计基于 Beta 校准的概率修正模

型，通过拟合 SEMs 预测结果的分布特性，建立置信度与真实准确率之间的可靠映射，从而为后续推理提供更为准确的可靠性指标。

(2) 设计错误驱动的原则生成算法。基于采样扩展的样例数据，针对 SEMs 预测错误的样本，逐步提炼可迁移的推理准则，形成一组指导任务推理的动态原则，帮助 LLMs 在 SEMs 发生错误时，也能作出精准的推理。

(3) 在涵盖新闻文本分类、生物医学问答、社会道德分类、电影情感分类的四个基准数据集上开展系统性实验。实验结果表明，CCDC 在总体准确率以及 F1 值上均有显著提升，并在不同上下文配置下展现出较高鲁棒性。

3.2　CCDC 框架

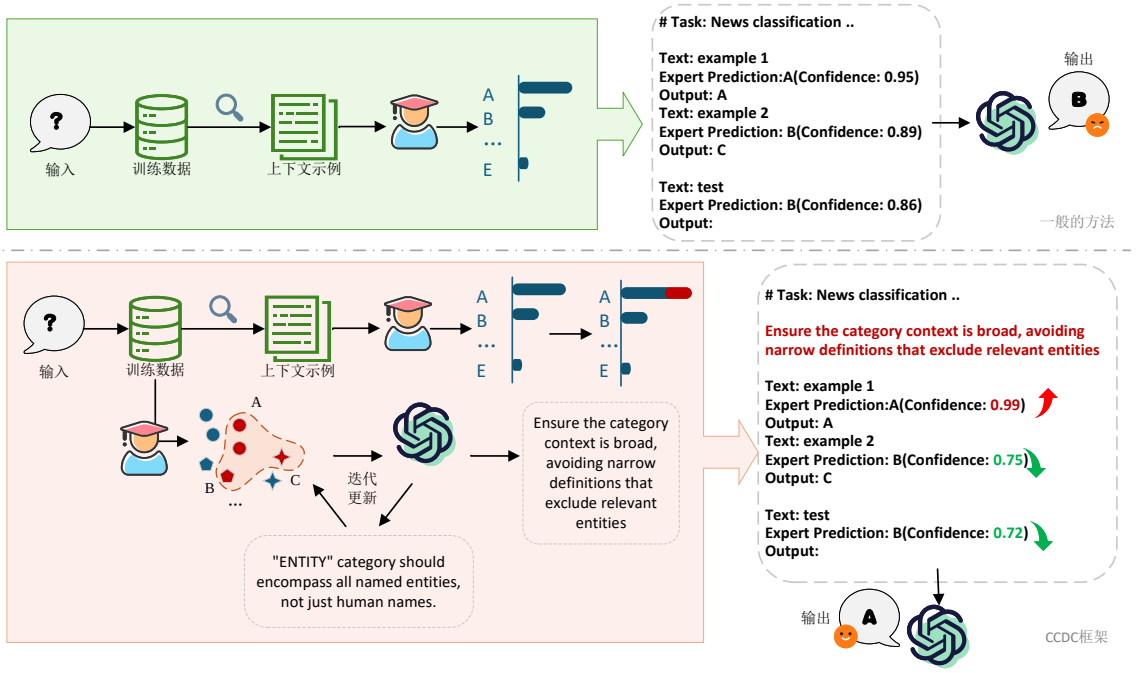


图 3-1　已有工作框架（上）和 CCDC 框架（下）对比

Fig. 3-1 Comparative Analysis of Existing Work (Top) vs. CCDC (Bottom)

整体框架如图3-1所示。图中上半部分展示了现有工作的典型架构，这些方法通过直接利用 SEMs 的预测结果作为辅助信号，将 SEMs 的输出与手工标注信息联合嵌入提示模板中，以实现对 LLMs 的指导；而下半部分则展示了本研究提出的 CCDC 框架。相较于传统方法，CCDC 在保留 SEMs 有效信息的同时，通过独立构建校准模型，实现对 SEMs 输出的置信度进行动态调整，并进一步通过错误分析模块提取易错样例中的推理准则，以在低置信度情景下更有效地引导 LLMs 优化推理结果。该框架整体流程分为三个阶段：置信度校准、错误驱动的动态原则生成以及动态决策。这三个阶段的协同作用旨在降低 SEMs 的固有偏差对整体系统

性能的负面影响，同时在低置信度时提升 LLMs 对复杂领域任务中的推理准确性。

3.2.1 置信度校准

在直接采用 SEMs 的输出概率作为置信度的指标时，由于模型自身存在的偏差，会导致概率评估失准。具体而言，部分预测结果存在概率虚高现象，而另一些结果则表现出低估倾向。这种偏差将直接影响 LLMs 对 SEMs 输出的可靠性评估。为此，本研究提出了一种基于概率校准的置信度优化方法。通过为每个类别独立构建校准模型，在保持原始预测结果不变的前提下，对输出的置信度进行动态调整，以降低 SEMs 固有偏差对 LLMs 置信度评估的干扰。

(1) 问题定义

给定包含 N 个样本的数据集 $D=\{x_i,y_i\}_{i=1}^{N}$，所有类构成集合 $C=\{1,2,\ldots,M\}$。对于任意输入样本 $x_i \in D$，$\forall m \in C$，SEMs 输出的类别概率分布可表示为：

$$p_i^{(m)} = \frac{e^{z_i^{(m)}}}{\sum_{k\in C} e^{z_i^{(k)}}} \tag{3-1}$$

其中，$z_i^{(m)}$ 为 SEMs 在类别 m 上未归一化的的原始逻辑值。

然而，这种未经校准的概率输出 $p_i^{(m)}$ 往往与样本的真实置信度存在系统性偏差，导致 SEMs 预测概率不能准确反映实际正确概率。因此，需要建立从原始概率空间到校准置信度空间的映射函数 $F:[0,1]^M \to [0,1]^M$。该函数旨在对每个类别的预测概率进行重新映射，使得校准后的概率更贴近真实的正确率分布，从而为后续推理过程中置信度的判定和错误驱动机制提供更为可靠的依据。

(2) Beta 校准算法

针对 SEMs 输出概率的特点，本研究采用 Beta 校准方法对概率空间进行动态调整。与传统的 Platt Scaling 等线性校准方法相比，Beta 校准具备两项核心优势：首先，它能够有效处理极端概率值（即接近 0 或 1 的预测），这对于 SEMs 表现出尖锐概率分布的情况尤为重要；其次，该方法通过灵活调整 Beta 分布的形状参数，可更好地适应不同数据分布，从而降低过拟合风险。对于二分类场景，Beta 校准函数定义为：

$$B(p;\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \tag{3-2}$$

其中，$B(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$ 为 Beta 函数，$\alpha>0$ 和 $\beta>0$ 为可学习的形状参数。

为了适应多分类场景，本研究工作采用 ‘‘一对一’’ 校准策略。具体而言，对于每个类别 $m \in C$，仅利用被 SEMs 预测为该类别的样本来训练对应的校准器 $B_m(\cdot)$。

设定每个类别 m 对应的训练子集为：

$$D_m = \{x_i \mid \arg\max_m p_i^{(m)} = m\} \tag{3-3}$$

针对每个子集 D_m，根据公式3-2求解各类校准模型 B_m 的参数 (α_m, β_m)。

(3) 校准策略实施

完成参数学习后，对于每个样本 x_i 和类别 m，实施校准策略。在校准过程中，遵循预测不变性原则，该原则体现在两个层次的要求中。首先，从数学形式上保证，SEMs 的硬标签预测结果 y' 始终保持不变，即保证原始的分类决策边界不受校准过程的影响；其次，在概率分布层面，仅允许对预测类别对应的置信度水平 $p_i^{(m)}$ 进行数值修正，而其他非预测类别的概率值保持不变。

具体而言，对于任意样本 x_i，其预测类别定义为：

$$m \triangleq y' = \arg\max_{k \in C} \mathbf{p}_i^{(k)} \tag{3-4}$$

校准后的输出概率分布采用条件式规则进行定义：当且仅当类别 k 为当前预测类别 m 时，对其置信度施加校准函数 $B_k(\cdot)$ 的转换操作，而对于其他类别，直接保留原始预测概率，即：

$$p_i^{(k)} = \begin{cases} B_k(p_i^{(k)}) & \text{当} m = k \\ p_i^{(k)} & \text{其他情况} \end{cases} \tag{3-5}$$

这种约束设计的必要性来源于两个维度的考量：

1) 在技术实现层面，硬标签的修正涉及决策边界的调整，本质上属于模型再训练或增强学习的范畴，与概率校准的原始目标存在本质差异；

2) 在理论逻辑层面，置信度校准的核心诉求在于建立概率估计值与真实正确率的一致性关系，而非提升模型本身的判别性能。

预测不变性约束一方面可以通过局部调整预测类别的置信度来修正系统偏差，另一方面还可以避免因修改预测结果可能引入的新误差源。整体流程见算法1。

3.2.2 错误驱动的动态原则生成

当 LLMs 接收到高置信度的预测结果时，通常可直接采纳 SEMs 的输出。然而在 SEMs 容易发生错误的低置信度情况下，缺乏相应的指导机制以辅助其做出更准确的判断。为解决该问题，本研究引入错误驱动的原则生成模块，从 SEMs 的错误预测中提取指导原则。该过程可分为三个阶段：

(1) 错误样本池构建

算法 1 基于 Beta 校准的置信度评估算法

输入： SEMs 输出的概率分布 $\{p_i^{(m)}\}_{i=1}^N$，对应样本标签 $\{y_i\}_{i=1}^N$，类别集合 C

输出： 校准后的置信度得分 $\{s_i^{(m)}\}_{i=1}^N$

初始化每个类别 $m \in C$ 的校准模型参数 (α_m, β_m)
for 每个类别 $m \in C$ **do**
    构建训练子集 $D_m = \{(p_i^{(m)}, \mathbb{1}_{[y_i=m]}) \mid \arg\max_k p_i^{(k)} = m\}$
    基于 D_m 拟合 Beta 校准模型 $B_m(p; \alpha_m, \beta_m)$
end for
for 每个样本 x_i **do**
    设预测类别 $m = \arg\max_k p_i^{(k)}$
    for 每个类别 $k \in C$ **do**
        if $k = m$ **then**
            $s_i^{(k)} = B_k(p_i^{(k)}; \alpha_k, \beta_k)$　　▷ 仅对预测类别施加校准
        else
            $s_i^{(k)} = p_i^{(k)}$　　▷ 其余类别概率保持不变
        end if
    end for
end for
return 校准后的置信度分布 $\{s_i^{(m)}\}_{i=1}^N$

在数据池 D 中筛选预测错误的样本集合：

$$D_{\text{error}} = \{(x_j, y_j, y'_j) \in D \mid y'_j \neq y_j\} \tag{3-6}$$

其中，y'_j 为 SEMs 的预测结果。

为保证原则生成过程中能够覆盖各类别的错误情况，还需根据类别分布采取分层采样策略，从 D_{error} 中抽取出 k 条具有代表性的错误样本。

(2) 原则种子生成

随机选取初始样本 $(x_0, y_0, y'_0) \in D_{\text{error}}$，并利用 LLMs 生成初始的错误指导原则，记为：

$$p_0 = \text{LLM}(\Psi_{\text{origin}}(x_0, y_0, y'_0)) \tag{3-7}$$

其中，Ψ_{origin} 为表6-1定义的提示模版，其目的是从初始错误案例中提取出有助于修正推理流程的种子原则。

(3) 原则渐进优化

采用迭代方法对初始种子原则进行不断优化，以增强原则的泛化性和适用性。对于第 t 次迭代，通过给定当前采样错误样本 x_t 以及相应更新提示 $\Psi_{\text{update}}(x_t)$(表6-

1)，生成更新后的原则：

$$p_t = \mathrm{LLM}(\Psi_{\mathrm{update}}(x_t, y_t, y_t')) \tag{3-8}$$

当预先设定的 k 条代表性错误样本均被遍历后，最终生成任务专属的动态原则 p_{k-1}，该原则将用于指导 LLMs 在低置信度场景下改进推理结果。整体算法见2。

3.2.3 动态决策

在引入推理准则的流程中，本研究构建了‘‘信任优先、修正补充’’的决策范式。经实证发现，经过概率校准后，SEMs 在高置信度区域的预测准确率显著高于低置信度区域。如图3-5-b) 所示，当置信度大于 0.75 时，其预测准确率均超过了预期平均水平。若对高置信结果进行无差别修正，不仅会增加计算成本，而且可能因修正规则的不完备性，干扰本已正确的预测。为系统性地识别 SEMs 认知不确定的区域，本研究采用统计方法设定动态阈值，从而为后续推理准则的启用提供精准触发条件。

具体而言，本研究采用置信度统计的方式，对预测结果按照其置信度进行排序，并从中选取后 25% 的置信度作为动态阈值，记为 τ。在实际应用中，对于任一输入样本，如果 SEMs 对某类别的预测置信度满足 $p_i^{(m)} > \tau$（即处于高置信度区域），则直接采纳该预测结果，无需额外引入修正准则，以最大限度地保证决策效率；而当 $p_i^{(m)} \leq \tau$ 时，则认为该预测存在较大的不确定性，此时系统将自动激活错误驱动模块，通过生成对应的推理准则，填充到预设模板（表6-2）中，指导 LLMs 进一步推导答案：

$$output_i = \begin{cases} \Psi_{\mathrm{generate}}(x_i, y_i', y_i) & \text{当} p_i^m > \tau \\ \Psi_{\mathrm{generate}}(x_i, y_i', y_i, p_{k-1}) & \text{当} p_i^m \leq \tau \end{cases} \tag{3-9}$$

通过这样的动态决策机制，可以在保障整体决策准确性的同时，有效补救低置信度预测带来的风险，提高推理系统在复杂场景下的鲁棒性。

3.3 实验与分析

为系统评估 CCDC 框架的有效性并深入探究其性能机理，本研究设计了广泛的实验评估框架，重点围绕以下三个核心研究问题展开：

(1) 与现有协同推理方法相比，CCDC 在整体推理性能上具有怎样的优势？

(2) 在不同上下文配置下，CCDC 的鲁棒性如何？

(3) CCDC 框架中各功能模块的作用与贡献程度如何？

算法 2 错误驱动的动态原则生成算法

输入: 数据集 D，SEMs 预测概率 $p_c(x)$，提示模板 $\Psi_{\text{origin}}, \Psi_{\text{update}}$，大语言模型 LLM，采样数量 k
输出: 最终生成的原则 p_{k-1}
初始化错误样本池：$D_{\text{error}} \leftarrow \{(x_j, y_j, y'_j) \in D \mid \arg\max_c p_c(x_j) \neq y_j\}$
按类别均衡策略从 D_{error} 中分层采样 k 条样本：$\{(x_0, y_0, y'_0), \ldots, (x_{k-1}, y_{k-1}, y'_{k-1})\}$
生成初始原则：$p_0 \leftarrow \text{LLM}(\Psi_{\text{origin}}(x_0, y_0, y'_0))$
for $t = 1$ **to** $k-1$ **do**
    $p_t \leftarrow \text{LLM}(\Psi_{\text{update}}(x_t, y_t, y'_t, p_{t-1}))$
end for
return p_{k-1}

表 3-1　各类数据集信息
Table 3-1 Information for Each Dataset

数据集	领域	# 训练集 A	# 训练集 B	# 测试集	# 类别
Trec	新闻	2 726	2 726	500	6
PubmedQA	生物医学	5 000	5 000	890	2
SST-5	电影	4 822	4 822	2 210	5
Prosocial Dialog	社会规范	5 000	5 000	1 500	5

3.3.1 基础配置

(1) 数据集与评估指标

为全面验证 CCDC 协同推理框架的跨任务泛化能力和鲁棒性，本研究构建了覆盖多种任务形态的四类基准数据集，涵盖文本分类、垂直领域推理、细粒度情感分析及社会价值观对齐，具体数据集及其描述如下：

- TREC：细粒度问题分类数据集，包含 5 952 个新闻领域的问题样本，划分为 6 个语义类别，主要用于评估模型在语义解析和问题分类任务中的表现。
- PubmedQA：生物医学问答数据集，包含基于 PubMed 摘要构建的超过 20 000 对 QA 样本。数据集领域术语密度高，重点考察模型在生物医学领域的科学推理能力和领域知识运用。
- SST-5：情感分析数据集，基于斯坦福情感树库的五分类扩展版本，包含 11 855 条电影评论。通过细粒度情感标签体系，从 ‘‘非常消极’’ 到 ‘‘非常积极’’，构建连续情感空间，对模型对语言细节的敏感性提出了更高要求。
- Prosocial Dialog：道德对话数据集，包含超过十万条多轮对话样本。该数据集模拟包含价值观冲突的社会交互场景，旨在评估模型在社会规范对齐方面的稳定性与鲁棒性。

上述四类数据集覆盖了从开放域到专业领域、从单句分类到多句对话等多种任务形态，为评估 LLMs 在不同领域与任务中的表现提供了丰富的基准。此外，在验证集中，SEMs 的准确率分布范围在 0.51 至 0.96 之间，形成阶梯式的性能分布，便于深入分析不同 SEMs 能力水平在 LLMs 协同效应中的具体表现。

为保证实验的可行性和执行时间，对于数据量较大的 PubmedQA 和 Prosocial Dialog 数据集，本研究分别进行了采样处理。表3-1详细列出了各数据集在采样后的分布情况，以及各数据集在不同任务指标上的初步统计结果。

评估指标方面，鉴于主要任务类型为分类问题，因此本研究采用准确率（Accuracy）和宏 F1 值（Macro-F1）。具体计算如下：

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i = \hat{y}_i) \tag{3-10}$$

$$\text{Macro}-\text{F1} = \frac{1}{C}\sum_{c=1}^{c}\frac{2\times P_c \times R_c}{P_c + R_c} \tag{3-11}$$

其中，N 为样本总数，y_i 为真实标签，$\hat{y}_i$ 为预测标签。C 为类别总数，$P_c = \frac{TP_c}{TP_c+FP_c}$ 为类别 c 的精确率，$R_c = \frac{TP_c}{TP_c+FN_c}$ 为召回率。

(2) 模型

SEMs 侧，本研究通过差异化的微调策略构建领域能力不同的 SEMs。

- 对于 TREC、Prosocial Dialog 和 SST-5 数据集，采用 RoBERTa-base[1]作为基础模型，在模型顶部添加单层分类器（隐藏层维度 → 类别数）进行任务适配。优化过程采用 AdamW 算法，批次规模为 64，通过端到端微调实现参数优化。
- PubmedQA 数据基于 BioLinkBERT-base[2]模型实施领域自适应微调，该模型通过文档链接感知预训练机制强化了跨文本知识关联能力，有效捕获生物医学文献的深层语义关联。

LLMs 侧，本研究选取三种具有代表性的开源 LLMs，分别是 GLM-4-9B-Chat[3]、Qwen-7B-Chat[4]和 Meta-LLaMA-3-8B-Instruct[5]。

- GLM-4-9B-Chat 是智谱 AI 推出的基于 GLMBlock架构 的多语言 LLMs，在语义理解、长文本推理等任务上均有出色表现。
- Qwen-7B-Chat 由阿里云团队研发，是一个专为中英文双语环境设计的聊天 LLMs。在 C-Eval、MMLU、HumanEval 等标准基准测试中表现优秀，展现出较高的准确率和鲁棒性。

[1] https://github.com/pytorch/fairseq/tree/master/examples/roberta

[2] https://github.com/michiyasunaga/LinkBERT

[3] https://huggingface.co/THUDM/glm-4-9b-chat

[4] https://huggingface.co/Qwen/Qwen-7B-Chat

[5] https://huggingface.co/meta-LLaMA/Meta-LLaMA-3-8B-Instruct

- Meta-LLaMA-3-8B-Instruct 是 Meta 基于 LLaMA 3 系列发布的指令调优 LLMs，主要面向英文对话与文本生成任务。

为确保结果的可复现性，所有 LLMs 推理过程均采用确定性解码策略，即设置 temperature=0，其他参数均采用默认配置。

(3) 运行环境

计算平台使用 2 张 NVIDIA RTX A4000 显卡（16GB 显存）。为确保实验可复现性，本研究为所有涉及的随机过程（包括 NumPy、PyTorch 等）设置了全局随机种子。所有 LLMs 推理均采用 BFloat16 精度计算。

(4) 基线与参数设置

为开展对比实验，本研究选取了以下基线方法：

- ICL：最基础的领域内推理方法，通过检索少量的样例作为上下文指导 LLMs 生成答案。
- SuperICL：在 ICL 的基础上，将 SEMs 的预测结果与上下文原始信息一并嵌入提示，以指导 LLMs 完成当前任务。
- SuperContext：在 SuperICL 的基础上，补充思维链，进一步整合上下文中其他样本的预测与标签信息，以增强 LLMs 的推理能力。

参数配置方面，考虑到 PubmedQA 的测试输入文本较长，因此 PubmedQA 的上下文数量为 0，其他数据集统一设置为 4。

3.3.2 对比实验及分析

(1) CCDC 框架在多个数据集上的表现显著优于现有协同推理基线方法。如表3-2所示，高亮表示性能优于最优对比基线的情况。CCDC（默认 Beta 校准）在多个数据集上的表现显著优于现有协同推理基线方法，主要体现于以下两个方面：

首先，CCDC 展现出较强的 LLMs 兼容能力。当以 GLM-4 作为基模型时，CCDC 在 TREC、PubmedQA 与 SST-5 三个数据集的所有评估指标上均取得最优结果；以 Qwen 为基模型时，在 7 项指标中达到最优，另有 1 项保持次优水平；以 LLaMA-3 为基模型时，尽管在 TREC 数据集上略逊于基线方法，但在 PubmedQA 与 SST-5 上仍表现出明显优势。这表明 CCDC 框架能够在不同架构的 LLMs 上稳定发挥作用，具备良好的通用性。其次，在 SEMs 的超越性方面，实验结果显示，基于 GLM-4 的 CCDC 在 SST-5 数据集上在各指标上均超越 SEMs 2.89 与 1.55 个百分点。在 LLaMA-3 为基模型的配置下，CCDC 在 SST-5 任务中首次实现 0.04 个百分点的反超。本研究认为，这种性能超越来源于基础 LLMs 在架构演进与训练语料规模上的提升。当 LLMs 能力达到一定阈值后，借助 CCDC 的协同机制，能够有效突破传统 SEMs 的性能边界。

表 3-2　随机采样下不同基准 LLMs 在协同框架下的跨数据集性能比较

Table 3-2 Cross-dataset Performance Comparison of Different Models under the Collaborative Framework with Random Sampling

模型	方法	Trec		PubmedQA		SST-5		Prosocial Dialog	
		ACC	F1-Score	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score
SEM	None	0.9660	0.9395	0.8461	0.8288	0.5344	0.5262	0.5130	0.4591
Qwen-7B-Chat	ICL	0.5860	0.5542	0.8090	0.7938	0.4633	0.3179	0.2227	0.1236
	SuperContext	0.7960	0.8085	0.7787	0.7483	0.4801	0.4475	0.3240	0.2941
	SuperICL	0.9080	0.9026	0.8281	0.8149	0.5262	**0.4860**	0.4633	0.3480
	CCDC w/ Beta	0.9140	0.9204	**0.8371**	**0.822**	**0.5285**	0.4855	**0.4793**	**0.3597**
	CCDC w/ Temp	**0.9218**	0.9032	0.8180	0.8017	0.5186	0.4687	0.4720	0.3429
	CCDC w/ Platt	0.9180	**0.9218**	0.8236	0.8079	0.5199	0.4647	0.4693	0.3421
GLM-4-9B-Chat	ICL	0.8240	0.7896	0.8034	0.7975	0.5217	0.4202	0.4553	0.3122
	SuperContext	0.9440	0.9020	0.8416	0.8310	0.5348	0.5115	0.4433	0.3809
	SuperICL	0.9580	0.9426*	0.8573*	0.8491	0.5570	0.5407	0.5053	0.4291
	CCDC w/ Beta	**0.9620**	**0.9461***	**0.8629***	**0.8548**	0.5633	0.5417	0.5113	**0.4319**
	CCDC w/ Temp	0.9600	0.9440*	0.8618*	0.8537	**0.5710**	**0.5480**	0.5113	0.4237
	CCDC w/ Platt	0.9600	0.9440*	0.8607*	0.8526	0.5688	0.5461	**0.5120**	0.4217
Meta-LLaMA-3-8B-Instruct	ICL	0.6660	0.6573	0.8225	0.8161	0.4833	0.4727	0.3680	0.3047
	SuperContext	**0.9240**	**0.9029**	0.8528	0.8394	0.5262	0.5016	0.4053	**0.3994**
	SuperICL	0.9040	0.8965	0.8674*	0.8597	0.5303	0.5158	**0.4393**	0.3944
	CCDC w/ Beta	0.9020	0.8951	**0.8742***	**0.8679**	**0.5348**	**0.5177**	0.4280	0.3940
	CCDC w/ Temp	0.9120	0.9008	0.8742*	0.8675	0.5240	0.5110	0.4153	0.3806
	CCDC w/ Platt	0.9040	0.8875	0.8742*	0.8674	0.5235	0.5110	0.4173	0.3753

[1] 最优与次优结果分别以**加粗**与下划线 表示。

[2] * 表示性能超越 SEMs。

[3] 表示 CCDC 性能超过最优基线。

(2) 选择 Beta 概率校准算法的正确性。除了 Beta 校准外，本文还对比了其他主流概率校准算法：温度校准（Temperature Scaling）以及 Platt 缩放（Platt scaling）校准。结果显示，基于 Beta 校准的 CCDC 具有更稳定的性能，在各 LLM 和各数据集上都有着超越基线的表现，而另外的概率校准算法则在 LLM 和数据集上均体现出较大的性能差异表现，体现了选择 Beta 概率校准的正确性。

3.3.3　鲁棒性实验及分析

为评估 CCDC 在不同配置下的鲁棒性，本研究采用控制变量实验设计，以分类准确率作为核心评估指标，考察 SEMs 的选择、检索策略与上下文数量对模型性能的影响。

(1) SEMs 选择

由于不同 SEMs 的泛化能力存在差异，可能影响实验结论的普适性。为验证 CCDC 方法在多种 SEMs 配置下的鲁棒性，本文除默认采用 BERT 模型外，还引

表 3-3　随机采样下使用不同 SEMs 的各方法在 TREC 数据集下的性能比较

Table 3-3 Performance Comparison of Various Methods Using Different SEMs under TREC Dataset with Random Sampling

模型	方法	ACC	F1-Score	模型	方法	ACC	F1-Score
Ernie	None	0.9640	0.9384	Bi-LSTM	None	0.8500	0.8594
Qwen-7B-Chat	ICL	0.5680	0.5650	Qwen-7B-Chat	ICL	0.5680	0.5650
	SuperContext	0.6860	0.7136		SuperContext	0.6220	0.6596
	SuperICL	0.8940	0.9028		SuperICL	0.8060	0.8041
	CCDC	**0.9200**	**0.9219**		CCDC	**0.8220**	**0.8122**

表 3-4　BM25 采样下不同基准 LLMs 在协同框架下的跨数据集 ACC 性能比较

Table 3-4 Cross-Dataset ACC Performance Comparison of Different Benchmark Models under the Collaborative Framework with BM25 Sampling

模型	方法	Trec		SST-5		Prosocial Dialog	
		ACC	F1-Score	ACC	F1-Score	ACC	F1-Score
SEM	None	0.9660	0.9395	0.5344	0.5262	0.5130	0.4591
Qwen-7B-Chat	ICL	0.6680	0.6314	0.4573	0.3511	0.3253	0.1834
	SuperContext	0.6720	0.7180	0.4679	0.4314	0.3200	0.2915
	SuperICL	0.9180	0.9079	0.5348*	0.4958	0.4747	0.3466
	CCDC	**0.9260**	**0.9172**	**0.5376***	**0.4994**	**0.4827**	**0.3506**
GLM-4-9B-Chat	ICL	0.9400	0.9174	0.5267	0.4479	0.4707	0.3099
	SuperContext	0.8320	0.8125	0.5403*	0.5173	0.4213	0.3618
	SuperICL	0.9620	0.9278	0.5557*	0.5391*	0.4887	0.3959
	CCDC	**0.9620**	**0.9330**	**0.5652***	**0.5426***	**0.5007**	**0.4048**
Meta-LLaMA-3-8B-Instruct	ICL	0.8660	0.8635	0.4937	0.4799	0.4013	0.3351
	SuperContext	0.8860	0.8546	**0.5321**	0.5014	0.3700	0.3799
	SuperICL	0.9420	**0.9155**	0.5249	0.5084	**0.4340**	**0.3940**
	CCDC	**0.9420**	0.9107	0.5294	**0.5112**	0.4267	0.3842

入了 Ernie-2.0-Base-En[6] 以及基于 GloVe[7] 向量的 Bi-LSTM 模型。

表3-3展示了在 TREC 数据集上，各 SEM 配置下不同方法的性能比较。结果表明，无论选用哪种 SEM，CCDC 始终保持最佳表现，进一步证明了其优越性结论的普适性。

(2) 检索方式

在检索方式上，除默认的随机上下文检索外，本研究引入了基于相似度的 BM25 检索策略进行对比分析。BM25 是一种经典的信息检索方法，其通过词频统计与文本长度归一化，对上下文之间的相关性进行量化评估，从而筛选出更具代表性的上下文示例。对于 PubmedQA 而言，因采用零样本推理，故不作比较。

[6] https://huggingface.co/nghuyong/ernie-2.0-base-en

[7] https://huggingface.co/DylanJHJ/glove.6B.300d

实验结果如表3-4所示。采用 BM25 策略后，CCDC 框架在多个数据集上均展现出显著性能提升：相较于 ICL 基线方法，准确率普遍提升了 5-10 个百分点；相较于次优方法 SuperICL，普遍提升 1 个百分点，进一步验证了 CCDC 在高相关性上下文支持下的协同推理能力。

值得注意的是，当基模型为 LLaMA-3 时，CCDC 与 SuperICL 之间的性能差距有所缩小。本研究推测其原因在于：一方面，LLaMA-3 所采用的稀疏注意力机制在处理高相关性输入时具备更强的模式捕捉能力，从而弱化了推理规则模块的边际增益；另一方面，BM25 筛选出的高相似度示例本身已蕴含充足的判别性信息，使得 SEMs 预测的固有噪声被显著压制，进而减少了校准模块的干预空间。但综合多场景测试结果，CCDC 仍保持显著的性能优势，证明其设计具备良好的性能。

尽管在部分配置下性能差距缩小，CCDC 在不同检索方式和多种任务背景下依然保持稳定且显著的性能优势，充分体现了其在多上下文配置下的鲁棒性。

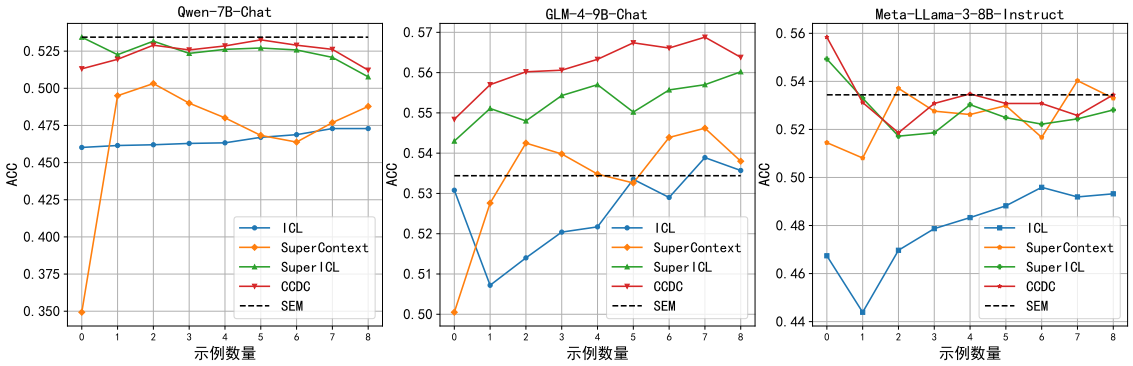


图 3-2　不同 LLMs 在 SST-5 数据集上的上下文示例数量敏感性分析

Fig. 3-2 Sensitivity Analysis of Contextual Example Quantity for Various Large Language Models on the SST-5 Dataset

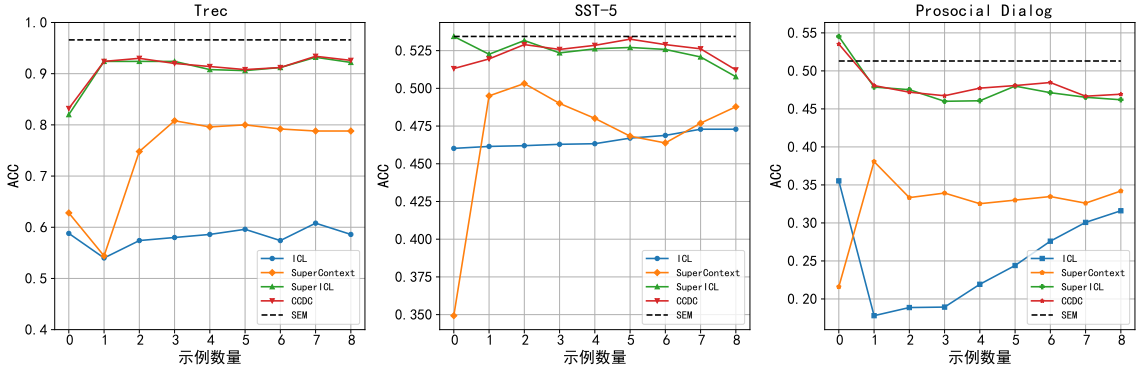


图 3-3　Qwen 跨数据集上下文示例数量适应性分析

Fig. 3-3 Adaptability Analysis of Qwen Across Datasets on Contextual Example Quantity

(3) 上下文数量

本研究还系统对比了上下文示例数量对 CCDC 的影响。

表 3-5　Qwen 在 Trec 和 PubmedQA 上的消融实验（颜色标注性能变化）

Table 3-5 Ablation Study Results of Qwen with Color-Coded Performance Variations on Trec and PubmedQA

方法	Trec		PubmedQA	
	ACC	F1-Score	ACC	F1-Score
CCDC	0.9140	0.9204	0.8371	0.8220
w/o 原则生成	0.9160(+0.20%↑)	0.9178(-0.26%↓)	0.8202(-1.69%↓)	0.8094(-1.26%↓)
w/o 概率校准	0.9120(-0.20%↓)	0.9186(-0.20%↓)	0.8371(-0.00%↓)	0.8211(-0.09%↓)
w/o 阈值过滤	0.9100(-0.40%↓)	0.9165(-0.39%↓)	0.8258(-1.13%↓)	0.8118(-1.02%↓)

表 3-6　Qwen 在 SST-5 和 Procial Dialog 上的消融实验（颜色标注性能变化）

Table 3-6 Ablation Study Results of Qwen with Color-Coded Performance Variations on SST-5 and Procial Dialog

方法	SST-5		Procial Dialog	
	ACC	F1-Sscore	ACC	F1-Score
CCDC	0.5285	0.4855	0.4793	0.3597
w/o 原则生成	0.5267(-0.18%↓)	0.4880(+0.25%↑)	0.4647(-1.46%↓)	0.3434(-1.63%↓)
w/o 概率校准	0.5262(-0.23%↓)	0.4855(-0.00%↓)	0.4733(-0.60%↓)	0.3552(-0.45%↓)
w/o 阈值过滤	0.5285(-0.00%↓)	0.4837(-0.18%↓)	0.4740(-0.53%↓)	0.3513(-0.84%↓)

从 LLMs 的适应性角度来看，如图3-2所示，在 SST-5 数据集上，不同 LLMs 对上下文示例数量的响应表现出一定差异。基于 GLM-4 的 CCDC 在所有上下文配置下均保持显著优势，显示出良好的性能稳定性。随着上下文数量增加至 3 个，Qwen 在 CCDC 框架下虽然性能提升幅度相较 GLM-4 略小，但仍展现出持续的性能增益。而 LLaMA-3 的性能随上下文数量变化呈现出非单调波动：在零样本配置下的性能优于少样本配置。结合前文3.3.3节的分析，本研究认为这可能与 LLaMA-3 固有的稀疏注意力机制有关。但随着示例数量增至 3 个，CCDC 的准确率相较 SuperICL 提升超过 1 个百分点，进一步验证了在足够上下文支持下，CCDC 能够有效引导 LLMs 进行更深层次的推理。

从任务多样性角度来看，图3-3展示了 Qwen 在不同数据集上的表现。实验表明，CCDC 在大多数测试场景中均保持最优性能，体现出良好的泛化能力。尽管在上下文示例较少（如 1-2 个）时，CCDC 与其他方法间的性能差距相对有限，但当示例数量增加至 4-6 个时，CCDC 的性能增益达到峰值，显示其在复杂上下文配置下的协同推理优势。

联合分析图3-2和图3-3可以发现，CCDC 对上下文规模的敏感性显著低于基线方法。以 Qwen 为例，其在 $k \geq 2$ 时性能曲线近似线性，而 GLM-4 与 LLaMA-3 虽未呈完全线性趋势，但在 CCDC 方法下的曲线波动较小，整体表现更为平稳。这一结果表明，CCDC 在中等规模（3-4 个）上下文支持下即可实现稳定的性能提升，无需依赖大量示例支撑。

综合来看，CCDC 在 LLMs 适应性、任务泛化性以及上下文数量敏感性方面均展现出显著优势，证明了该框架在多种上下文配置下的高鲁棒性与实用价值。

3.3.4　消融实验及分析

本研究基于 Qwen 对 CCDC 框架进行了全面的消融实验，主要对比了以下三种设置：

(1) 不进行概率校准：即移除概率校准模块；

(2) 不采用错误驱动的原则生成：即不引入从错误中提取生成的指导原则；

(3) 取消在推理时对原则设置动态阈值的机制。

此外，本研究还绘制了 LLaMA-3 在 SST-5 数据集上（图3-5），不同置信度区间下的准确率分布图，以便直观展示置信度调整对 LLMs 表现的影响。

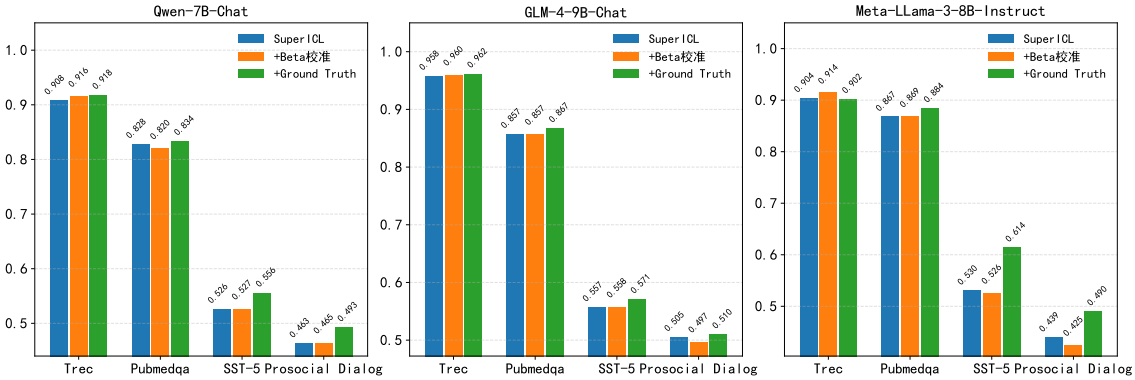
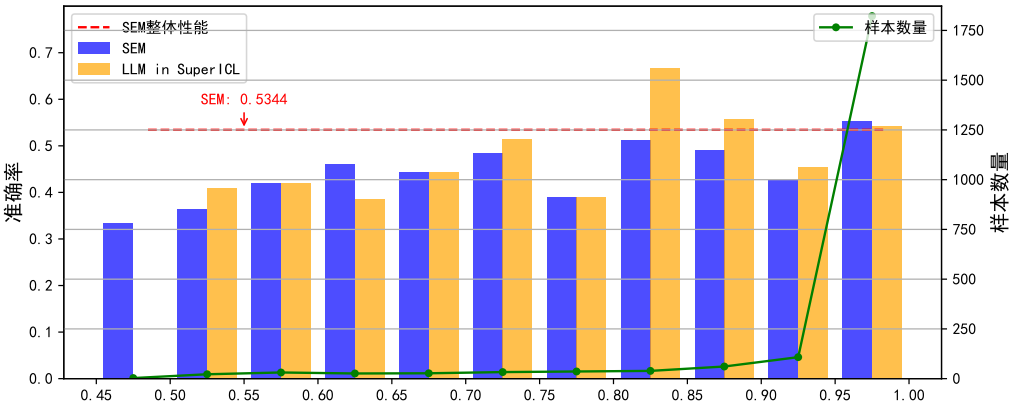


图 3-4　概率校准有效性分析

Fig. 3-4 Analysis of Probability Calibration Effectiveness

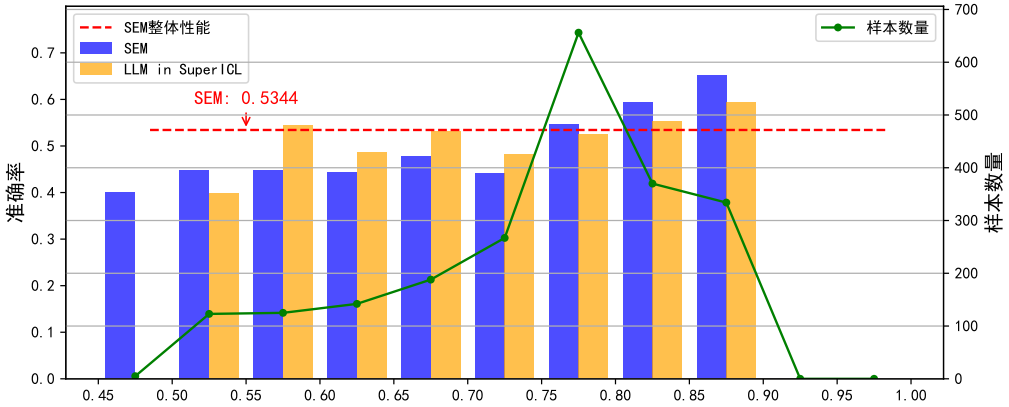
(1) 在不进行概率校准的情况下，实验结果（表3-5和表3-6 ）表明，在四个数据集上，准确率和 F1 值均出现不同程度的下降。从图3-5-a) 也可以观察到，未经过校准的 SEMs 预测结果在整体上存在较大的波动性，而图3-5-b) 中的校准后的预测结果相对更为稳定。随着置信度提升，校准后的区间内准确率逐渐上升，符合本研究的预期：即在 SEMs 输出更高可信的结果时，其预测结果也应更为可靠，LLMs 在高置信度情况下的预测准确性也得以稳定。但也可以发现，当校正后的置信度较低时，LLMs 的性能存在一定下降。

(2) 在移除错误驱动的原则生成模块的实验中，多个数据集均出现了不同程度的指标下降。结合图3-5-c) 与图3-5-b) 的对比，可以看出两者主要区别在于低置信度区间 LLMs 的预测准确性。图3-5-c) 中引入了错误原则信息，使得系统在低置信度区间能够获得额外推理指导，从而显著提升了这一部分的性能；而图3-5-b) 中缺少该模块，导致 LLMs 在低置信度情景下的表现较为逊色。这一结果验证了错误驱动原则生成模块在提升低置信度预测准确性方面的有效性。



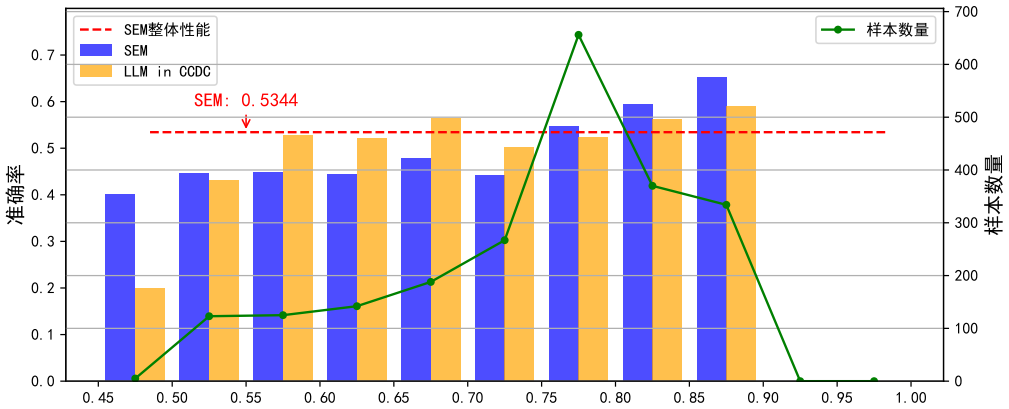
a) 概率校正前各方法的性能分布图

a) Performance Distribution Plots for Each Method before Probability Calibration



b) 概率校正后各方法的性能分布图

b) Performance Distribution Plots for Each Method after Probability Calibration



c) CCDC 与 SEMs 的性能分布图

c) Performance Distribution Plots of CCDC and SEMs

图 3-5　不同方法下的置信度区间与准确率的关系

Fig. 3-5 Relationship Between Confidence Intervals and Accuracy Across Different Methods

(3) 在取消动态阈值设置的情况下，表3-5和表3-6的实验结果显示性能进一步下降。取消阈值机制后，使得 LLMs 在整个推理过程中对所有样本一视同仁地应用修正准则，这不仅使得高置信度情况下产生额外的不必要修正，也削弱了错误驱动模块在低置信度区域引导 LLMs 推理的效果。该实验结果说明动态阈值机制在区分并精确触发规则修正方面具有重要作用，从而显著保障了整体框架的性能。

(4) 本研究还对概率校准模块的天花板性能进行了分析，即探究在概率预测完全准确的理想情况下，CCDC 框架能够达到怎样的性能。具体而言，本研究对上下文示例和测试示例的概率进行了如下处理：如果 SEMs 预测错误，则将预测概率设置为 $1/c$，其中，$c = |C|$ 是类别的数量；如果 SEMs 预测正确，那么预测概率设置为 1。图3-4展示了基线方法、加入 Beta 校准模块后的性能以及校准天花板情况下的性能比较。结果显示，当预测概率达到理想状态时，LLMs 性能有了显著提升，三者呈现阶梯分布，这从侧面验证了概率校准模块的正确性和潜在提升能力。

3.4 本章小结

本章围绕有标签场景下的大小模型协同，聚焦两个关键问题：一是 SEMs 原始预测概率存在系统性偏差，使得其置信度评估结果与实际准确率不符；二是在 SEMs 预测错误时，没有对应的优化机制。为此，本研究提出了基于概率校准与错误驱动的有标签协同推理算法。首先，本研究基于 Beta 校准对 SEMs 输出的概率进行动态修正，有效处理置信度与实际准确率之间的不一致问题。其次，本研究设计了一套错误驱动的原则生成流程。该流程从预测错误的样本中进行采样，经过迭代优化形成可迁移的推理准则。这一机制使得 LLMs 在 SEMs 发生错误时，能够借助从错误中提炼出的领域知识进行更为深层次的推理。最后，通过在多个基准数据集上的广泛实验，结果表明 CCDC 在准确率及宏 F1 值上均获得了显著提升，并在不同上下文配置下展现出较强的鲁棒性。

4 基于多源知识验证的无标签协同推理算法

4.1 引言

尽管在有真实标签的场景中，可以通过联合利用 SEMs 提供的弱标签信号与人工标注的真实标签，以提升 LLMs 在垂直领域的推理能力，但现实中获取高质量标注往往代价高昂，构建难度较大。因此，也有研究开始转向无真实标签条件下，探索 LLMs 与 SEMs 之间的协同推理机制。

在无真实标签场景下，融合 SEMs 的 LLMs 的推理增强策略主要沿着知识表征的泛化性改进展开。这种思路将 SEMs 的标签信息扩展为偏好信息，将预测的偏好转换成更具有泛化性的语言描述的知识，以检索增强的形式提升 LLMs 的领域适应性[10,11]。然而，这类方法存在一定的局部信息局限性。这种局限性主要体现在两个方面：其一是知识粒度的局限性。在知识表征层面，现有方法通常将知识生成限制于单个实例级别，忽略了对领域整体知识结构的建模，难以捕捉跨任务、跨实例的高阶语义特征；其二是知识源的局限性。在知识溯源层面，若简单采信 SEMs 的偏好性输出作为知识生成依据，可能因模型固有问题导致错误知识的传播。

针对知识表征的局限性，分层协同机制在知识工程领域已被证明具备显著优势[67]。例如，在分层文本分类[68] 以及基于分层标签抽取的研究中[69]，分层结构被广泛用于提升知识整合的深度与准确性。近期的一项研究[66] 进一步将层次聚类方法引入 LLMs 推理任务，构建了任务级与实例级的双层错误诊断框架，并在七个基准数据集上取得了显著性能提升，为本研究构建多层次知识验证体系提供了有力的实践支撑。

与此同时，在另一类聚焦于模型规模差异的知识蒸馏框架中，弱监督驱动的大小模型协同同样展现出较强的泛化能力。OpenAI 团队提出的“弱到强”范式[55] 即是典型代表：他们采用较小规模语言模型（弱学生）生成的标签作为监督信号，引导更大规模的语言模型（强学生）进行上下文学习，从而实现更强的推理性能。然而，该方法受到理论性能边界的限制，即强学生的性能上限取决于弱学生的监督强度，难以突破经典蒸馏框架中的“教师瓶颈”问题[70]。

为突破上述限制，研究者们提出了一类针对弱监督信号优化的新方法，代表性做法是“可扩展监督”（Scalable Oversight）[71]，旨在借助 AI 辅助标注手段提高信号的可靠性与多样性。例如，近期基于辩论机制的多角度对抗框架[72,73] 通过整合异质预测视角，显著提升了监督信号的鲁棒性与泛化能力。这种多源验证思路

也为本研究提升第二个局限性—— SEMs 标注质量提供了新的方向。

综合上述两类方法的讨论，本研究提出一种基于多源知识验证的无标签协同推理算法（Dual-Source Knowledge Integration, Dual-Know），旨在更好地利用 SEMs 的偏好信息以提升知识的准确性与知识表征的语义一致性。该算法通过全局-局部协同策略缓解局部知识局限性，并引入多源辩论机制优化局部知识生成流程，从而实现兼具领域广泛性与细粒度知识精度的无标签推理增强。

因此，本章研究工作的主要贡献如下：

(1) 构建全局-局部协同的双粒度知识验证体系。针对单粒度知识易产生认知偏差的问题，提出分层协同策略：在全局层面，利用分层聚类算法构建语义一致的领域聚类簇，基于簇内样本，形成全局的领域共性知识表征；在局部层面，通过单个样本分析提取局部知识。全局与局部知识进行双向校验，有效缓解认知偏差；

(2) 设计基于多假设辩论的局部共识知识生成算法。突破传统方法直接采纳 SEMs 首优预测结果的局限，提出多候选序列辩论决策机制：首先在 SEMs 的原始预测上进行调整，生成优先次序不同的候选假设，为每个序列生成对应知识陈述；然后构建假设评估模块，由 LLMs 筛选更有说服力的局部知识作为最终输出，降低对单一 SEMs 决策路径的依赖性；

(3) 在三个标准数据集上的实验表明，Dual-Know 在准确率与 F1 值指标上均显著优于基准方法。通过控制训练数据规模的扩展实验进一步证明，Dual-Know 在低资源场景下仍保持稳定的性能表现，验证了算法的强鲁棒性。

4.2 Dual-Know 框架

本节提出一种基于多源知识验证的无标签协同推理算法 Dual-Know，旨在提升 LLMs 在无标签场景下的推理准确性与鲁棒性。该算法通过语义聚类驱动的全局知识，构建多粒度、多来源的知识体系，并引入多假设偏好的知识选择与验证机制，缓解局部偏见与噪声干扰带来的推理误导问题。

方法整体由三个核心模块构成：(1) 基于语义聚类驱动的全局知识生成模块，挖掘语义相似样本间的共性模式，产出鲁棒的簇级领域知识；(2) 构建多假设的局部知识生成机制，通过对 SEMs 的多个候选预测进行语言层面的知识建模与对抗评估，生成更具说服力的局部知识；(3) 双粒度知识推理机制，将全局-局部双层级的知识表征融入提示中，实现交叉验证，从而提升整体推理结果的可靠性和鲁棒性。整体框架如图4-1。

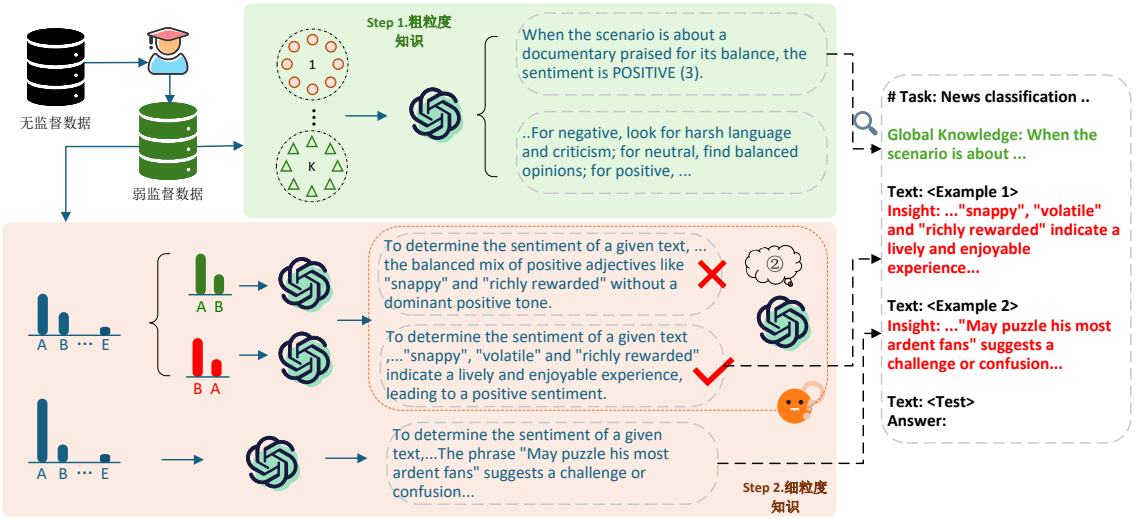


图 4-1 Dual-Know 的算法框架

Fig. 4-1 The Framework of Dual-Know

4.2.1 基于聚类的全局知识生成

全局领域知识基于多源数据生成，具备较广的覆盖范围。通过对语义表示相近的问题样本进行聚类，能够挖掘跨样本的共性知识模式。一方面，簇内样本在隐空间中的相似性表明其潜在知识需求具有一致性，有助于降低单样本局部知识可能引发的过拟合风险；另一方面，聚类生成的全局知识可为局部知识提供跨实例的验证支持，从而提升 LLMs 决策能力。

(1) 语义驱动的聚类建模

具体而言，首先使用语义向量模型 E 将无标签数据集 D 中每个样本 x_i 映射为语义表示向量 $h_i \in \mathbb{R}^d$，其中，d 为隐空间维度。基于此构建样本语义空间 $H = \{h_i\}_{i=1}^N$，并采用 $K-means$ 算法将其划分为 k 个簇：

$$C = \{C_j\}_{j=1}^k = \underset{\{\mu_j\}_{j=1}^k}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{h_i \in C_j} \|h_i - \mu_j\|^2 \tag{4-1}$$

其中，$\mu_j \in \mathbb{R}^d$ 表示第 j 个簇的质心，$\|\cdot\|$ 为欧氏距离度量。通过优化该目标函数，语义相近的样本被划分至同一簇 C_j，而质心 μ_j 则表征该簇的语义原型。

(2) 全局知识生成

对于每个簇 C_j，选取与质心 μ_j 距离最近的 m 个样本组成代表性子集 $S_j = \{x^{(l)}\}_{l=1}^m$，以捕捉簇内的典型语义特征。该子集的选取满足以下优化目标：

$$S_j = \arg \min_{S \subseteq C_j,\ |S|=m} \sum_{x_l \in S} \|h_l - \mu_j\|_2^2 \tag{4-2}$$

其中，$\|\cdot\|_2^2$ 表示欧氏距离，h_l 为样本 x_l 的语义表示向量。代表性子集 S_j 将用于驱动 LLMs 生成簇级全局知识。

算法 3 基于聚类的全局知识生成算法

输入: 无标签数据集 $D=\{x_i\}_{i=1}^{N}$，语义编码器 E，聚类数 k，代表子集大小 m，提示模板 Ψ_{cluster}
输出: 全局知识集合 $K_{\text{global}}=\{k_{\text{global}}^{(j)}\}_{j=1}^{k}$
对每个样本 x_i 计算语义向量 $h_i=E(x_i)$
构建语义空间 $H=\{h_i\}_{i=1}^{N}$
对 H 执行 K-means 聚类，得到簇集合 $C=\{C_j\}_{j=1}^{k}$ 及质心 μ_j ▷ 语义驱动的聚类建模
for $j=1$ to k **do**
    从 C_j 中选择与 μ_j 最接近的 m 个样本构成代表子集 S_j
    构造结构化提示 $prompt_j=\Psi_{\text{cluster}}(S_j)$
    $k_{\text{global}}^{(j)}=\text{LLM}(prompt_j)$ ▷ 全局知识生成
    $K_{\text{global}}\leftarrow K_{\text{global}}\cup\{k_{\text{global}}^{(j)}\}$
end for
return K_{global} ▷ 返回完整的全局知识集合

定义 1 (全局知识). 给定代表性子集 S_j，通过预设模板 $\Psi_{cluster}$（见表6-3）组织为结构化提示，驱动 *LLMs* 生成簇级全局知识：

$$k_{global}^{(j)}=LLM(\Psi_{cluster}(S_j)) \tag{4-3}$$

整体流程见算法3。

4.2.2 基于多假设的局部知识生成

直接基于 SEMs 的预测偏好生成局部知识，容易受到其自身偏好的影响，可能导致在 LLMs 生成带有幻觉的局部知识，进而影响后续推理效果。为解决这一问题，本章提出一种基于多假设的局部知识生成算法。该方法引入候选偏好集合，并通过辩论机制引导 LLMs 做出更为准确的局部知识决策。

(1) 候选偏好生成

给定无标签数据集 $D=\{x_i\}_{i=1}^{N}$ 作为检索池，通过对小型专家模型系统 *SEM* 的概率计算，生成有序预测结果。*SEM* 的输出概率空间记为 $P(y|x_i)$，对于任意样本 $x_i\in D$，其有序预测序列可形式化表示为：

$$Y_i'=\text{sort\_desc}\left(P(y|x_i)\right),\quad y\in Y \tag{4-4}$$

其中，k 表示预测类别数，sort_desc$(\cdot)$ 表示按类别概率从高到低排序的操作，Y 为

标签空间，$Y_i' = (y_i^{(0)}, y_i^{(1)}, ..., y_i^{(k-1)})$ 表示按置信度降序排列的预测序列，用于后续的多假设生成与知识辩论，满足 $p(y_i^{(m)}|x_i) \geq p(y_i^{(n)}|x_i)$ 当且仅当 $m < n$。

为提取可泛化的局部决策知识，本研究使用基于偏好对的知识映射机制。该机制通过 SEMs 的预测标签揭示显式推理知识，同时从预测序列的排序关系中挖掘隐式偏好知识 。

定义 2 (基础偏好对)**.** *设 Y_i' 为样本 x_i 的预测序列，基础偏好对定义为有序二元组 $P_i^{base} = (y_i^{(0)}, y_i^{(1)})$，其中，首优预测 $y_i^{(0)}$ 与次优预测 $y_i^{(1)}$ 构成对比关系。通过提示模板 (见表6-3) 驱动 LLMs 进行知识解析。*

在基础偏好对构建完成后，为了获得多辩手观点，需要在基础偏好对的基础上扩展形成候选偏好集。该扩展策略可以显著提升潜在正确预测的覆盖范围。

有工作[74] 发现 SEMs 和 LLMs 在置信度不同的情况下，SEMs 相比 LLMs 在置信度较低的困难样本（硬样本）上表现不佳，而置信度较高的简单样本（软样本）上表现更好，因此，本研究使用置信度阈值区分软硬样本。对于软样本，候选偏好对即基础偏好对；对于硬样本，候选偏好对来自扩展的不同偏好对的集合。

定义 3 (候选偏好对)**.** *令 $Acc@t$ 表示 SEMs 的 $top-t$ 预测准确率，当存在扩展系数 $\alpha \geq 1$ 使得 $Acc@2 \geq \alpha Acc@1$ 时，可通过置换与跨阶操作生成扩展偏好对：*

$$P_i^{ext} = \bigcup_{0 \leq m < n \leq t-1} \left\{ (y_i^{(m)}, y_i^{(n)}), (y_i^{(n)}, y_i^{(m)}) \right\} \tag{4-5}$$

$$|P_i^{ext}| = 2 \cdot \binom{t}{2} = t(t-1) \tag{4-6}$$

当考虑 $top-2$ 预测时，候选对数量扩展为 $|P_i^{\text{ext}}| = 2$，包含原始对及其置换对 $\{(y_i^{(0)}, y_i^{(1)}), (y_i^{(1)}, y_i^{(0)})\}$。

(2) 候选论点生成

候选偏好集合 P_i^{ext} 中的每个有序偏好对 $(y_i^{(m)}, y_i^{(n)})$ 对应辩论中的一个知识主张 $k_{(m,n)}$。所有生成的知识主张构成评估的论域 K。在决策阶段，这些主张将作为 LLMs 的候选论点，支持其进行多视角综合判断。

具体实现时，将每个偏好对按预设模板构建提示语句，输入 LLMs 后生成对应的论证知识：

$$k_{(m,n)} = \text{LLM}\left(\Psi_{\text{argue}}(y_i^{(m)}, y_i^{(n)})\right) \tag{4-7}$$

其中，Ψ_{argue} 为表6-3定义的论辩提示模板，K_i 则构成了样本 x_i 的决策论域：

$$K_i = \left\{ k_{(m,n)} \mid \forall (y_i^{(m)}, y_i^{(n)}) \in P_i^{\text{ext}} \right\} \tag{4-8}$$

(3) 多假设决策

多假设决策基于候选知识集合 K_i 进行知识择优推理。

为了帮助LLMs能够更好的完成决策,算法基于 x_i 对4.2.1节得到的全局知识进行检索,得到与 x_i 最相近的全局知识 $k^{(j)}_{\text{global}}$。然后通过提示工程将 $K_i=\left\{k_i^1,\ldots,k_i^{t(t-1)}\right\}$ 中的多视角观点，按照预设的评估模板进行结构化组织，形成包含输入问题 x_i，候选知识及评估准则的提示。判官 LLM 基于其大规模预训练获得的语义理解与逻辑推理能力，通过对比分析各候选知识的适配性，输出综合相比更正确的候选知识作为局部知识 $k^{(i)}_{\text{local}}$:

定义 4 (最优局部知识)**.** 给定样本 x_i 及其候选知识集合 K_i，局部知识定义为满足最大适配准则的候选知识:

$$k^{(i)}_{local}=LLM\left(\Psi_{judge}(x_i,K_i,k^{(j)}_{global})\right) \tag{4-9}$$

其中，Ψ_{judge} 为表6-3定义的论辩提示模板。

整体流程见算法4。

4.2.3 全局-局部的双粒度知识推理

在决策阶段，给定新样本 x_i，通过双粒度知识协同实现稳健推理。该机制通过全局知识保证跨样本的一致性约束，同时利用局部知识捕捉特定实例的细粒度特征，形成互补增强的推理范式。

(1) 全局知识检索

首先进行全局知识的语义匹配。通语义向量模型 E 将样本 x_i 映射为语义向量: h_i。然后计算 h_i 与所有簇质心 $\{\mu_j\}_{j=1}^{k}$ 的余弦相似度，确定最近邻簇:

$$j^*=\arg\max_{j=1}^{k}\ \frac{h_i^\top \mu_j}{\|h_i\|\cdot\|\mu_j\|} \tag{4-10}$$

最后检索定义1中的全局知识集合得到全局知识 $k^{(j^*)}_{\text{global}}$。该过程建立样本与语义原型的关联，为后续推理提供领域共性知识锚点。

(2) 局部知识检索

为捕捉实例特异性，本研究进一步检索局部知识。计算 x_i 与无标签数据集 D 中所有样本的语义相似度:

$$sim_l^{(i)}=\frac{h_i^\top h_l}{\|h_i\|\cdot\|h_l\|},\quad \forall x_l\in D. \tag{4-11}$$

h_i 和 h_l 分别是数据的语义向量。选取相似度最高的 k 个样本构成邻域集合:

$$N_i=\left\{x_l\in D\,\middle|\,\text{rank}_{\downarrow}(sim_l^{(i)})\le k\right\} \tag{4-12}$$

算法 4 基于多假设的局部知识生成算法

输入: 无标签数据集 $D=\{x_i\}_{i=1}^{N}$，专家模型 SEM，大语言模型 LLM，置信度阈值 τ
输出: 局部知识集合 $K_{\text{local}}=\{k_{\text{local}}^{(i)}\}_{i=1}^{N}$
for 每个样本 $x_i \in D$ **do**
  获取预测概率序列 $P(y|x_i)$，按置信度降序排列得到 $Y_i'=(y_i^{(0)},y_i^{(1)},...,y_i^{(k-1)})$
  构造基础偏好对 $P_i^{\text{base}}=(y_i^{(0)},y_i^{(1)})$
  if $P(y_i^{(0)}|x_i)<\tau$ **then** ▷ 若为硬样本，扩展候选偏好对
    构造扩展偏好集 $P_i^{\text{ext}}=\{(y_i^{(m)},y_i^{(n)}),(y_i^{(n)},y_i^{(m)})\}_{0\le m<n<t}$
  else
    $P_i^{\text{ext}} \leftarrow \{P_i^{\text{base}}\}$
  end if
  初始化候选知识集 $K_i \leftarrow \emptyset$
  for 每个候选偏好对 $(y_i^{(m)},y_i^{(n)}) \in P_i^{\text{ext}}$ **do**
    构造提示 $\Psi_{\text{argue}}(y_i^{(m)},y_i^{(n)})$
    生成论点 $k_{(m,n)}=LLM(\Psi_{\text{argue}})$
    $K_i \leftarrow K_i \cup \{k_{(m,n)}\}$
  end for
  检索全局知识 $k_{\text{global}}^{(j)}$，满足 $x_i \approx x^{(j)} \in S$
  构造决策提示 $\Psi_{\text{judge}}(x_i,K_i,k_{\text{global}}^{(j)})$
  获取最优局部知识 $k_{\text{local}}^{(i)}=LLM(\Psi_{\text{judge}})$
  $K_{\text{local}} \leftarrow K_{\text{local}} \cup \{k_{\text{local}}^{(i)}\}$
end for
return K_{local}

提取对应样本的局部知识集合 $\mathbf{k}_{\text{local}}^{(i)}=\left\{k_{\text{local}}^{(l)}\right\}_{x_l \in N_i}$。其中，$k_{\text{local}}^{(l)}$ 由定义4生成，反映邻域样本的个性化决策依据。

(3) 双粒度推理机制

将全局与局部知识通过结构化模板6-4融合，驱动 LLMs 进行综合推理：

$$\hat{y}_i=\text{LLM}\left(\Psi_{\text{fuse}}(x_i,k_{\text{global}}^{(j^*)},\mathbf{k}_{\text{local}}^{(i)})\right) \tag{4-13}$$

整体算法流程如5所示。

算法 5 全局-局部的双粒度知识推理

输入: 数据集 $D = \{x_i\}_{i=1}^{N}$，簇质心集合 $\{\mu_j\}_{j=1}^{k}$
输出: 推理结果集合 $\{\hat{y}_i\}_{i=1}^{N}$
for 每个样本 $x_i \in D$ **do**
    语义向量$h_i \leftarrow E(x_i)$
    $j^* \leftarrow \arg\max_{j=1}^{k} \frac{h_i^\top \mu_j}{\|h_i\| \cdot \|\mu_j\|}$
    $k_{\text{global}}^{(j^*)} \leftarrow$ 全局知识检索(j^*) ▷ 检索与簇 j^* 相关的全局知识
    for 每个样本 $x_l \in D$ **do**
        $sim_l^{(i)} \leftarrow \frac{h_i^\top h_l}{\|h_i\| \cdot \|h_l\|}$
    end for
    $N_i \leftarrow \text{rank}_{\downarrow}(sim_l^{(i)}) \leq k$
    $\mathbf{k}_{\text{local}}^{(i)} \leftarrow \{k_{\text{local}}^{(l)}\}_{x_l \in N_i}$ ▷ 提取邻域样本的局部知识
    $\hat{y}_i \leftarrow \text{LLM}\left(\Psi_{\text{fuse}}(x_i, k_{\text{global}}^{(j^*)}, \mathbf{k}_{\text{local}}^{(i)})\right)$ ▷ 融合全局与局部知识，通过 LLMs 推理生成最终决策
    将 $\hat{y}_i$ 存储到推理结果集合 $\{\hat{y}_i\}_{i=1}^{N}$
end for
return $\{\hat{y}_i\}_{i=1}^{N}$ ▷ 返回所有样本的推理结果

4.3 实验与分析

本章节主要介绍了 Dual-Know 在不同场景下的表现情况，研究将围绕以下三个问题展开介绍：

(1) 与现有无标签协同推理方法相比，Dual-Know 在整体推理性能方面具有什么优势?
(2) Dual-Know 在不同配置下的鲁棒性如何?
(3) Dual-Know 框架中各功能模块的作用与贡献程度如何?

4.3.1 实验设置

(1) 数据集、评价指标与模型

在数据集选择方面，考虑到 PubmedQA 样本输入较长，实验延续第3章中的三个代表性数据集：TREC、PubmedQA 和 SST-5。

评估指标依旧采用 ACC 与 F1-Score 两种通用分类指标，分别衡量模型的整体预测能力与类别间的均衡性。

在模型设置方面，SEMs 部分，TREC、PubmedQA 和 SST-5 所采用的模型与

第3章保持一致。LLMs 部分，由于第3章中 Qwen 在推理任务中整体性能指标远低于 GLM-4 和 LLaMA-3，因此本章实验将其替换为性能更优的 Qwen2.5-7B-Instruct[1] 以进一步探究其推理效果。

(2) 对比基线

本研究设置以下四类基线方法，包括三种提示引导增强方法以及一种基于知识注入增强的协同方法，以系统评估 Dual-Know 的有效性：

- Few-shot：该方法基于 SEMs 生成的弱标签，构建上下文学习示例，模拟典型的弱到强泛化范式。其遵循经典的上下文学习范式，通过少量示例提示，引导 LLMs 完成推理任务。
- Few-shot+Conf：在 Few-shot 的基础上，增设置信度机制以补充第3章中有标签方法的置信度信息。该方法要求 LLMs 在进行推理时，同时评估预测结果的置信程度，从而进一步提升推理性能。
- Few-shot+CoT：在 Few-shot 框架基础上引入思维链推理机制，显式引导 LLMs 进行分步逻辑推导。具体做法是在输入中增加显式推理的要求，生成中间推理步骤，再输出最终答案，以提升推理的可解释性。
- Panda：该方法基于 SEMs 的偏好标注构建泛化性知识库，通过大小模型协同机制实现动态知识检索。SEMs 负责生成偏好标签，LLMs 则基于偏好输出进行见解生成，然后基于检索到的相关见解进行最终推理。

(3) 实验控制

为确保实验结果的可复现性与对比的公平性，本研究采用如下控制策略：

- 固定随机种子，设置 LLMs 的温度系数为 0，避免生成过程中的随机性对结果产生干扰；
- 所有方法统一使用语义检索策略，基于 paraphrase-MiniLM-L6-v2[2] 生成文本向量，并采用余弦相似度进行相似知识段或示例的检索，以保证输入信息的一致性与公平性。
- 平衡上下文示例数量与检索段落数量为 2 个，该设置经过预实验验证，既能保留足够的提示信息，又可避免输入长度超出 LLMs 处理能力导致的性能波动。

4.3.2 对比实验及分析

在默认配置下，Dual-Know 在所有任务中均展现出优异的整体推理性能。具体结果见表4-1。

[1] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[2] https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2

表 4-1 不同基座 LLMs 在不同方法下的跨数据集性能比较

Table 4-1 Cross-Dataset Performance Comparison of Large Language Models under Different Methods

模型	方法	Trec		SST-5		Prosocial Dialog	
		ACC	F1-Score	ACC	F1-Score	ACC	F1-Score
SEM	None	0.9660	0.9395	0.5344	0.5262	0.5130	0.4591
GLM-4-9B-Chat	Few-shot	0.8120	0.8298	0.5566	0.5269	0.4060	0.3736
	Few-shot+Conf	0.8080	0.8231	0.5484	0.5303	0.4193	0.3859
	Few-shot+CoT	0.7780	0.8097	0.5204	0.4756	0.3853	0.3655
	Panda	0.8560	0.8635	0.5367	0.5110	0.4060	0.3897
	Dual-Know	**0.8720**	**0.8834**	**0.5738**	**0.5460**	**0.4293**	**0.3902**
Qwen2.5-7B-Instruct	Few-shot	0.5460	0.5248	0.4371	0.4185	0.3773	0.3288
	Few-shot+Conf	0.5620	0.5180	0.4606	0.4333	0.3720	0.2819
	Few-shot+CoT	0.5400	0.5693	0.4529	0.4146	0.3600	0.3467
	Panda	0.8020	0.8115	0.5145	0.4907	0.4247	0.4012
	Dual-Know	**0.8340**	**0.8542**	**0.5421**	**0.5120**	**0.4387**	**0.4013**
Meta-LLaMA-3-8B-Instruct	Few-shot	0.6540	0.6647	0.3928	0.3865	0.3407	0.3201
	Few-shot+Conf	0.6320	0.6284	0.4213	0.4145	0.3380	0.3184
	Few-shot+CoT	0.6800	0.6609	0.4593	0.4515	0.3040	0.3016
	Panda	0.7240	0.7288	0.4647	0.4585	0.3607	0.3390
	Dual-Know	**0.7520**	**0.7345**	**0.4919**	**0.4940**	**0.3767**	**0.3501**

在没有真实标签的场景中，现有的提示增强方法面临明显挑战。以 TREC 数据集为例，GLM-4 和 LLaMA-3 在引入置信度信息或思维链提示后，模型表现不升反降，说明单纯依赖 SEMs 标签或链式推理等内部引导手段，难以稳定提升推理效果。为缓解上述问题，Panda 尝试结合 SEMs 偏好标签与检索增强机制，引入外部知识以促进 LLMs 泛化，但在不同任务和模型配置下仍存在性能波动。

与上述方法相比，Dual-Know 却在三个基座 LLMs 上均有着稳定领先的表现。例如，在 GLM-4 上，Dual-Know 在 Trec 任务中取得了 0.8720 的 ACC，显著高于 Panda（0.8560）与 Few-shot 方法（0.8120）；在 Prosocial Dialog 任务上，其 ACC 达到 0.4293，相较于 Panda 与 Few-shot 均有约 2.3 个百分点的提升。类似的趋势也出现在 Qwen2.5 与 LLaMA-3 上。这表明 Dual-Know 所提出的协同推理机制并不依赖特定的 LLMs 架构或参数规模，具有良好的通用性与适配能力。

Dual-Know 之所以能在多个模型和任务中持续取得优异表现，关键在于其对知识生成与验证流程的系统性设计。通过构造多组候选知识，并引入决策机制，Dual-Know 显著提升了知识的准确性与局部合理性。这种高度容错的处理策略使其在面对任务语义变动或模型风格差异时，依然能够保持稳定而高质量的推理输出。

表 4-2 使用不同 SEMs 的各方法在 TREC 数据集下的性能比较

Table 4-2 Performance Comparison of Various Methods Using Different SEMs under TREC Dataset

模型	方法	ACC	F1-Score	模型	方法	ACC	F1-Score
Ernie	None	0.9640	0.9384	Bi-LSTM	None	0.8500	0.8594
GLM-4-9B-Chat	Few-shot	0.8140	0.8406	GLM-4-9B-Chat	Few-shot	0.7660	0.8004
	Few-shot+Conf	0.8180	0.8315		Few-shot+Conf	0.7720	0.7969
	Few-shot+CoT	0.7980	0.8314		Few-shot+CoT	0.7620	0.8065
	Panda	0.8580	0.8642		Panda	0.8120	0.8234
	Dual-Know	**0.8600**	**0.8710**		Dual-Know	**0.8360**	**0.8427**

4.3.3 鲁棒性实验及分析

(1) SEMs 选择

为验证 CCDC 方法在多种 SEM 配置下的鲁棒性，避免不同 SEMs 的泛化能力的影响，本节依旧使用 Ernie-2.0-Base-En 以及基于 GloVe 向量的 Bi-LSTM 模型代替基础 SEM，以探究各方法的性能表现差异。

表4-2展示了在 TREC 数据集上，各 SEM 配置下不同方法的性能比较。结果表明，Dual-Know 不受 SEMs 变化的影响，始终保持最佳表现，证明了其优越性结论的普适性。

(2) 检索规模

为评估 Dual-Know 在不同检索条件下的鲁棒性，本节考察了其在检索规模从 100% 逐步缩小至 15% 时的性能变化，比较对象包括三个数据集上的主要基线方法，相关结果见图 4-2。

1) Dual-Know 能够良好适配不同检索规模。从整体表现来看，Dual-Know 的准确率曲线始终位于各基线方法之上，表现出稳定的领先优势。以 15% 检索规模为例，Dual-Know 在三个数据集上的准确率均较次优基线高出超过 2 个百分点，说明即使在低资源检索条件下，Dual-Know 依托其优化的检索策略与知识验证机制，也依然保持了较高性能，充分体现出其良好的适应性与灵活性。

2) Dual-Know 在不同检索规模下的性能波动幅度较小。表4-3展示了各数据集在不同检索规模下的性能统计，包括平均准确率、准确率方差以及最大差值情况。其中，平均准确率指标反映了整体性能水平，而准确率方差和最大差值则分别展示了不同检索规模条件下各方法的波动情况。从实验数据可以看出，Dual-Know 在 Trec 和 SST-5 数据集中分别取得了 0.0132 和 0.0003 的最小方差，这说明 Dual-Know 在面对从 15% 到 100% 这一广泛范围内的检索规模变化时，能够保持稳定且一致的性能输出。同时，最大差值指标也表明，不论是在极小还是极大量的检索条件下，Dual-Know 的性能波动均处于较低水平，其表现优于其他基线方法，反映出 Dual-Know 具有较高的鲁棒性。

表 4-3　检索规模实验下的统计特征

Table 4-3 Statistical Characteristics under Retrieval Ratio Experiments

数据集	方法	平均准确率↑	准确率方差↓	最大差值↓
Trec	Few-Shot	0.7989	0.0149	0.0420
	Few-Shot+Conf	0.7946	0.0158	0.0440
	Few-Shot+CoT	0.7800	0.0156	0.0540
	Panda	0.8423	0.0145	0.0480
	Dual-Know	**0.8597**	**0.0132**	**0.0420**
SST-5	Few-Shot	0.5465	0.0066	0.0194
	Few-Shot+Conf	0.5434	0.0076	0.0227
	Few-Shot+CoT	0.5060	0.0069	0.0231
	Panda	0.5458	0.0067	0.0226
	Dual-Know	**0.5729**	**0.0033**	**0.0104**
Prosocial Dialog	Few-Shot	0.4037	**0.0053**	**0.0147**
	Few-Shot+Conf	0.0110	0.4061	0.0326
	Few-Shot+CoT	0.4040	0.0122	0.0340
	Panda	0.3839	0.0107	0.0373
	Dual-Know	**0.4204**	0.0057	0.0193

Dual-Know 表现出的高适配性和低波动性主要源于以下两个方面：首先，全局知识通过语义聚类提取了跨样本的共性知识，这种高层次的知识表征提供了一种跨检索规模的一致性保障，使得 LLMs 在信息检索的覆盖范围变化时，依然能捕捉到稳定的语义锚点。其次，经过多假设辩论和候选知识优化的局部知识生成模块，使得 LLMs 能够在不同检索规模下，根据具体语义和实例特征灵活调整决策。这一过程不仅过滤了可能产生的噪声，更确保了在信息量较少或较多的场景下，局部决策的可靠性和一致性得到有效保障。

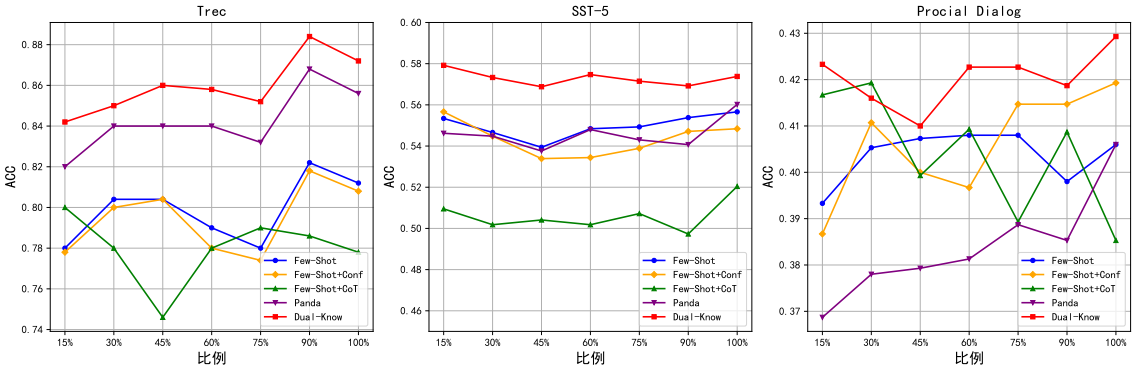


图 4-2　GLM-4 在不同数据集上的检索规模敏感性分析

Fig. 4-2 Sensitivity Analysis of GLM-4 to Retrieval Pool Ratios Across Datasets

(3) 检索数量

本研究进一步考察了不同检索数量对模型性能的影响。实验结果见图4-3，在三个数据集上，Dual-Know 均保持了显著的性能优势，展现出优于基线方法的稳

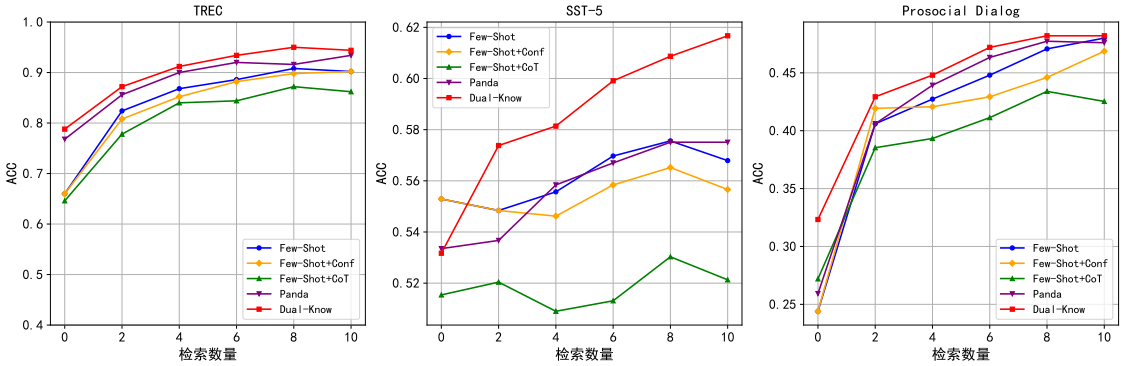


图 4-3　GLM-4 在不同数据集上的检索数量的影响分析

Fig. 4-3 Impact Analysis of Retrieval Example Quantities on GLM-4 Across Datasets

定性与泛化能力。

随着示例数量的增加，传统方法及 Panda 在性能上呈现出一定程度的波动。这种波动主要源于 LLMs 对输入数据分布变化的敏感性，尤其在示例数量较多时，冗余或不一致的信息可能对 LLMs 决策产生干扰，尤其体现在 SST-5 中。相比之下，Dual-Know 在示例数量逐步增加的过程中仍能持续提升准确率，说明该方法生成的知识相对稳健。一方面，全局知识从大量数据中提取出跨样本的共性知识，捕获了各类别的普遍特征，使得 LLMs 在面对不同示例数量时，依然能维持对总体语义趋势的准确把握；另一方面，经过校正的局部知识生成机制进一步提升了决策的精细化处理能力。因此在应对多样化输入时具备更强的鲁棒性。

4.3.4　消融实验及分析

为探究 Dual-Know 各组成部分的影响，本研究进行了消融实验，对以下三个模块进行了性能评估：(1) 全局知识；(2) 局部知识优化；(3) 局部知识决策阈值设置。实验结果如表4-4 所示，具体分析如下：

(1) 取消提示中全局知识的引入后，实验结果显示性能下降超过 1 个百分点，表明全局知识在 LLMs 推理过程中发挥了显著作用，有效促进了推理的正确性。这是因为全局知识是对大量数据进行聚类得到的高层次、抽象的知识表征。这种表征能够捕捉数据中的共性，提供比单个样本更全面的背景信息。同时，当全局知识被引入提示中时，LLMs 能够依托这种摘要知识，在处理边界模糊或信息冗余的情境时，避免因细节偏差而造成的决策错误。当取消全局知识后，模型性能下降，正是因为失去了这一稳定和整体视角的支持，导致在关键信息的捕捉和判断过程中出现偏差或误判，从而影响了推理结果的正确性。

需要指出的是，尽管全局知识在绝大多数场景中实现了 ACC 与 F1 的同步提升，但在 LLaMA-3 的 SST-5 与 Prosocial Dialog 两个任务的消融中，出现了 ACC

表 4-4 不同 LLMs 下的消融实验

Table 4-4 Ablation Study under Different LLMs

模型	方法			Trec		SST-5		Prosocial Dialog	
	全局知识	局部知识优化	局部决策阈值	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score
GLM-4-9B-Chat	✓	✓	✓	0.8720	0.8834	0.5738	0.5460	0.4293	0.3902
		✓	✓	0.8480(-2.40%↓)	0.8575(-2.59%↓)	0.5606(-1.32%↓)	0.5398(-0.62%↓)	0.4093(-2.00%↓)	0.3857(-0.45%↓)
	✓			0.8680(-0.40%↓)	0.8799(-0.35%↓)	0.5719(-0.19%↓)	0.5453(-0.07%↓)	0.4213(-0.80%↓)	0.3868(-0.34%↓)
		✓		0.8120(-6.00%↓)	0.8207(-6.27%↓)	0.5602(-1.36%↓)	0.5400(-0.60%↓)	0.3893(-4.00%↓)	0.3732(-1.70%↓)
Qwen2.5-7B-Instruct	✓	✓	✓	0.8340	0.8542	0.5421	0.5120	0.4387	0.4013
		✓	✓	0.8100(-2.40%↓)	0.8139(-4.03%↓)	0.5122(-2.99%↓)	0.4877(-2.43%↓)	0.4193(-1.94%↓)	0.3967(-0.46%↓)
	✓			0.8440(+1.00%↑)	0.8620(+0.78%↑)	0.5330(-0.91%↓)	0.5042(-0.78%↓)	0.4367(-0.20%↓)	0.3992(-0.21%↓)
		✓		0.7500(-8.40%↓)	0.7503(-10.39%↓)	0.5090(-3.31%↓)	0.4824(-2.96%↓)	0.4087(-3.00%↓)	0.3743(-2.70%↓)
Meta-LLaMA-3-8B-Instruct	✓	✓	✓	0.7520	0.7345	0.4919	0.4940	0.3767	0.3501
		✓	✓	0.7400(-1.20%↓)	0.7286(-0.59%↓)	0.4909(-0.10%↓)	0.4987(+0.47%↑)	0.3647(-1.20%↓)	0.3512(+0.11%↑)
	✓			0.6180(-13.40%↓)	0.5619(-17.26%↓)	0.4891(-0.28%↓)	0.4919(-0.21%↓)	0.3680(-0.87%↓)	0.3375(-1.26%↓)
		✓		0.7080(-4.40%↓)	0.6905(-4.40%↓)	0.4878(-0.41%↓)	0.4823(-1.17%↓)	0.3620(-1.47%↓)	0.3450(-0.51%↓)

下降而 F1 略有上升的情况。对此，本研究认为主要原因在于 Dual-Know 在全局知识生成与验证过程中强化了对主流类别的结构化抽象与一致性建模，从而有效提升了多数类的准确率。然而，对于样本量本就稀少的少数类别，其特征表征能力可能因信号覆盖不足而受到限制，导致召回率下降。虽然多数类性能的提升能够显著拉高整体 ACC，但在同时考虑精确率与召回率的 F1 指标下，少数类的性能下降会被更明显地放大，最终导致 ACC 上升而 F1 略降的指标差异。

(2) 若取消局部知识的决策阈值，即对所有示例一律采用局部知识进行决策优化，将带来不利影响。这体现在表4-4第三、七、十行。一方面，这导致局部知识生成过程显著延长，从而增加计算成本与响应时延，降低整体效率。更重要的是，这导致了决策准确性的下降。LLMs 在整体数据集上的表现往往不如专门训练的 SEMs。取消阈值后，LLMs 在大部分较为简单或 SEMs 已经能够准确判断的样本上也介入决策，反而可能引入额外的判断错误，导致整体性能下降，部分模型甚至出现了超过 5% 的性能下降。因此，设置阈值能够使 LLMs 仅在对 SEMs 来说具有一定挑战性的样本上发挥作用，保证决策的针对性和准确性。

(3) 此外，进一步取消局部知识的决策优化模块，仅依赖 SEMs 原始预测的偏好生成局部知识，也会导致性能下降。决策优化模块可以利用多种信息源和细粒度的判断机制，对候选答案进行筛选和优化，进而提高最终推理的准确性。取消这一机制后，LLMs 失去了对局部信息进行进一步加工和提升的机会，使得局部知识的描述不够全面和准确，从而导致整体性能下降。验证了通过扩大候选答案范围而获得更准确局部知识的决策优化模块在整体框架中的必要性。

4.4 本章小结

本章提出了一种基于多源知识验证的无标签协同推理算法，针对传统单一知识生成方式在知识局限性和认知偏差方面的局限性，设计了全局-局部协同的双粒度知识验证体系和基于多假设辩论的局部知识生成算法。具体而言，本研究在全局层面通过分层聚类提取了高层次、抽象化的领域共性知识，有效缓解了单样本知识提取中可能出现的认知局限；在局部层面，则通过生成多组候选局部知识并引入结构化的辩论决策机制，实现了对 SEMs 输出偏好的扩展与校正，从而获得更准确的局部知识表征。实验结果表明，所提出的方法在不同数据集和多种配置下均显示出优异的性能。在准确率与 F1 值指标上，Dual-Know 不仅显著优于传统的上下文学习、思维链和专家偏好方法，在检索规模和示例数量变化的复杂场景中也表现出较高的鲁棒性。对各模块的消融实验进一步验证了全局知识与局部知识双重设置的有效性。

5 面向自动报告生成的大小模型协同框架

为了验证大小模型协同在实际场景中的价值，本章选取自动报告生成这类复杂领域进行实际应用。自动报告生成是通过计算机算法自动提取原始数据或文本中的关键信息，按照既定格式和结构生成具有特定目的和内容的报告的过程。该技术已广泛应用于多个领域，例如医学影像报告[75]、社会金融报告[76]、企业财务报告[77]、信息检索报告[78] 等。自动报告生成的优势在于，它能够减轻人们在报告编写任务上的负担，节省阅读文献、分析数据和重复文本编写工作所需的大量时间和精力。此外，自动生成的报告具有一定的参考价值，能够为行业专家提供撰写报告的参考，甚至作为内容补充。

随着 LLMs 的出现，凭借其强大的语义理解和生成能力，有望在保持报告内容自然流畅的同时，提供更加详尽的解释和透明的推理路径，生成易于理解和信任的报告。一些工作探究了 LLMs 在自动报告生成上的性能。面对数据量较小的场景下，Ding[79] 直接将有限的原始文本输入到 LLMs 中，直接生成报告。这种方法简单直接，但真实报告场景下往往需要处理大量文档。另一些工作倾向于人为规定某些 SEMs 用于数据处理，如信息提取[80]、主题聚类[80] 等，LLMs 则根据处理后的数据进行指定内容的推理和生成。这种人为设计的半模板式 LLMs 报告生成方法利用 SEMs 在特定任务上的专业性和高效性，解决了 LLMs 的输入长度问题，但同时也存在一些局限性。

首先，这些方法并没有充分发挥 LLMs 在工具使用和自主规划方面的能力，在生成报告时依然需要依赖人工设定报告模块和对应的 SEMs 类型，这限制了报告内容的全面性和多样性。其次，这种输入提示并直接输出报告的工作流程会由于 SEMs 的错误或冗余的输出和 LLMs 自身生成过程的随机性而不稳定，导致部分报告的质量有所下降。

基于上述存在的两个问题，本章提出了一种面向自动报告生成的大小模型协同框架 AutoRG，旨在通过充分发挥 LLMs 与 SEMs 的互补优势，在确保报告质量的前提下，将人从报告生成的过程中解放（见图5-1）。AutoRG 框架将报告生成任务划分为“大纲规划与评估”、“工具调用与内容修正”和“报告生成与优化”三个阶段。“大纲规划与评估”阶段旨在通过 LLMs 的规划能力，定制化生成和筛选报告所需的子模块，确保每个子模块都能有合适的 SEMs 辅助完成；“工具调用与内容修正”阶段利用 LLMs 的工具理解能力选择性挑选 SEMs，结合优秀的摘要生成能力对 SEMs 输出进行修正，得到价值密度较高的分析数据；而“报告生成与优化”阶段旨在通过多智能体协作工作流的方式，迭代生成精确、可解释性强且具

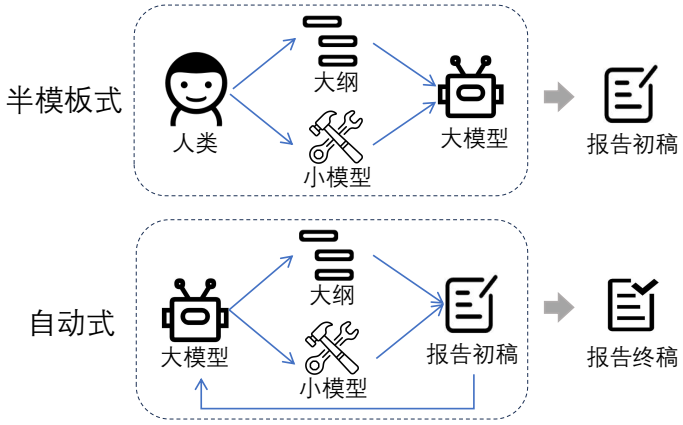


图 5-1 半模板式与自动式 LLMs 报告生成框架的对比

Fig. 5-1 Comparative Analysis of Semi-Templated vs. Fully-Automated LLM Reporting Frameworks

有深度的报告内容。总结来说，本章研究工作的主要贡献可以概括为：

(1) 本章提出了一个面向自动报告生成的大小模型协同框架 AutoRG，有效基于多个 SEMs 进行调用，自动化生成更灵活、全面的报告，无需过多的人工干预。

(2) 本章通过引入信息修正和报告迭代机制，降低对 SEMs 输出和 LLMs 随机性的依赖，提高了报告质量的稳定性。

(3) 本章以自动专利技术报告生成为例，采用主客观评估结合的方式，从多个维度验证了框架的有效性。

5.1 相关工作

传统自动报告生成方法主要分为抽取式和生成式两大类。

5.1.1 抽取式自动报告生成

抽取式报告生成方法主要通过自然语言处理的 SEMs，通过规则筛选[81]、主题聚类[80]、数据统计[78] 等技术手段直接从原始数据中提取相关信息并填充到预设的模板中。Babour[78] 提出一个自动生成指定国家科学计量研究分析报告的模型，通过使用数据统计 SEMs 对科学出版物数据进行详细的统计分析，填入模板并交由评审员进行评审修改得到分析报告。Noh[82] 提出一种名为 Wise Issue Report 的自动报告生成系统，利用主题生成 SEMs 生成不同主题，通过检索和多文档摘要为每组新闻生成摘要，结合时间事件摘要 SEMs 生成不同的摘要内容，组合得到完整报告。

抽取式报告的优势在于能够迅速获取基础信息，但其极度依赖模板，灵活性较差，难以适应多变的报告需求，同时也缺乏对内容进行深入分析的能力。

5.1.2 生成式自动报告生成

传统的生成式方法通过端到端训练的模式，如编码器-解码器模型，试图模仿人类编写报告的过程。这类方法通过学习原始报告中的潜在变量分布[83,84]，并结合知识图谱等先验知识[85,86]，从而生成更加自然和连贯的报告文本。Wang[87] 提出了一种基于条件变分自动编码器的方法，使用 Bi-GRU 编码输入的新闻，通过知识蒸馏和教师-学生网络结构来细化解码器组件的输出，实现金融报告的自动生成。Ren[76] 提出了一种新的混合深度生成神经模型，使用具有注意力机制的指针生成器网络学习输入新闻的大纲，通过改进的变分自动编码器生成宏观金融报告。Li[86] 为了提高放射学报告自动生成的质量，提出知识增强注入框架，通过结合医学概念和相似报告中的临床信息三元组，有效提高了报告的准确性和流畅性。

端到端训练的生成式方法虽然能够使模型掌握一定的报告语言规律并生成连贯的文本，但也往往导致模型过分关注于模仿训练数据中的表面特征，忽视深层语义逻辑。生成的报告可能在形式上看似合理，但在内容深度和逻辑性上却难以达到要求，可解释性不足。此外，高质量的领域报告资源相对稀缺，也限制了其对新领域或新主题报告的泛化能力。

5.1.3 基于 LLMs 的自动报告生成

LLMs 在自动报告生成领域中正逐渐成为一种趋势。得益于 LLMs 在语义理解与文本生成方面的强大能力，它为生成具有深度和广度的报告内容提供了新的可能性。这类方法结合抽取式和生成式报告生成的优点，采用分工合作的策略，其中 LLMs 主要负责生成连贯的文本内容，而 SEMs 则专注于原始数据的信息抽取，以辅助 LLMs 生成更为详尽和信息丰富的报告。Colverd[88] 提出了一个利用 LLMs 自动生成洪水灾害报告的系统，SEMs 负责检索关键词相似文档，LLMs 对文档进行评估并提取与洪水事件相关的信息，通过提示的方式让 LLMs 回答预设的问题，从而生成一个内容全面，结构完整的洪水灾害报告。Reddy[89] 利用 LLMs 自动生成态势报告，使用层次聚类进行新闻分类，利用 SEMs 生成类别标题，使用 GPT3 基于生成的类别标题进行战略子标题和内容的生成。曾等人[90] 融合信息抽取模型和 LLMs，生成全面且内容丰富的文档报告。

这些分工合作的方法虽然能够弥补传统生成式的不足，但仍然存在一些局限性。首先，生成报告的初步阶段通常需要人工设定报告的内容与使用的 SEMs，这可能会导致报告结构或内容的不全面。例如，在行业趋势预测中，除了行业标签和统计数据外，可能还需考虑龙头企业和专家的发展信息，以增加报告的丰富度和深度。此外，LLMs 在生成报告时可能会因为回复的不稳定性而产生质量波动。

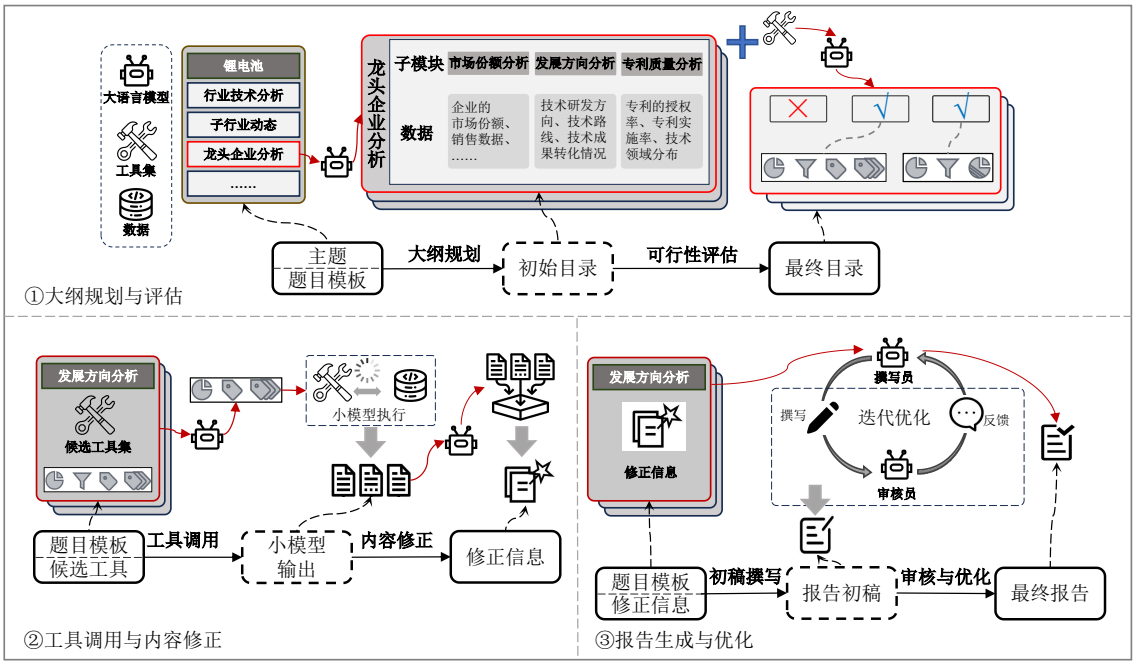


图 5-2　AutoRG 报告生成框架

Fig. 5-2 Framework of AutoRG

这种不稳定性一方面来自 SEMs 输出的错误与冗余信息，另一方面来自 LLMs 生成文本时的随机性。与人类不同，LLMs 缺乏自主修正和完善的能力，这也限制了报告内容的优化。

5.2　AutoRG 框架

在本节中，本章研究工作将详细介绍 AutoRG 框架，如图5-2所示。框架将报告生成的过程分为三个阶段。第一阶段，提供预设的一级题目和可利用的 SEMs 集合，利用 LLMs 生成尽可能丰富且可实现的子模块集合，将其作为报告的二级目录（第5.2.1节）；第二阶段，利用 LLMs 的工具理解与摘要生成的能力，为每个子模块生成高信息量且更精准的分析数据（第5.2.2节）；第三阶段，以多智能体协作的方式对报告内容进行迭代优化，获得最终的报告（第5.2.3节）。

5.2.1　大纲规划与评估

大纲规划阶段的目标是为报告创建一个详尽可实现的目录结构。算法流程如6所示。

(1) 大纲规划

首先，将预定义的题目模板集合 $T=\{T_1,T_2,\ldots,T_n\}$ 和可用的 SEMs 集合 $E=\{E_1,E_2,\ldots,E_m\}$ 交给 LLMs。通过精心设计的提示（表6-5），引导 LLMs 针对每个

题目模板 T_i 生成一个候选的子模块集合 $S_i = \{S_{i1}, S_{i2}, \ldots, S_{ik}\}$。这些子模块将构成报告的初始二级目录。在生成子模块 S_{ij} 时，要求 LLMs 不仅提供模块的标题，还需要预测可能需要的信息 D_{ij}，以便后续的数据收集，即：

$$S_i = \mathrm{LLM}_{\mathrm{gen}}(T_i, E) = \{(S_{ij}, D_{ij})\}_{j=1}^{k} \tag{5-1}$$

$$S_{ij} \rightarrow \mathrm{Title}_{ij} \tag{5-2}$$

(2) 可行性评估

尽管 LLMs 能够生成丰富的子模块集合，但并非所有生成的子模块都能通过现有的 SEMs 集合来实现。因此，需要对这些子模块进行筛选，确保它们与现有的 SEMs 集合相匹配。LLMs 需要对每个生成的子模块进行反思并做出两个关键决策：确定哪些子模块是可以执行的，以及哪些 SEMs 可能适用于这些子模块。这是 SEMs 首次被引入决策过程，目的是为了简化后续工具调用任务。

具体来说，设每个子模块 S_{ij} 对应的数据需求为 D_{ij}，小型专家模型 E_p 的能力描述为其可处理的数据类型集合 A_p。由 LLMs 评估每个子模块的可行性及匹配专家子集，则可表示为：

$$(\mathrm{flag}_{ij}, E_{ij}^{\mathrm{match}}) = \mathrm{LLM}_{\mathrm{eval}}(S_{ij}, D_{ij}, E, A) \tag{5-3}$$

其中，$\mathrm{flag}_{ij} \in 0, 1$ 表示子模块是否可执行，$E_{ij}^{\mathrm{match}} \subseteq E$ 表示可用于该子模块的 SEMs 集合。如果 $\mathrm{flag}_{ij} = 1$，则保留该子模块；否则，将其排除出最终大纲：

$$S_i^{\mathrm{final}} = \{(S_{ij}, D_{ij}) \in S_i \mid \mathrm{flag}_{ij} = 1\} \tag{5-4}$$

通过引入 LLMs 对执行能力的判断，使得生成模块不仅具有内容合理性，同时具备可实现性，从而为后续的专家模型调用和数据处理提供结构化保障。

5.2.2 工具调用与内容修正

工具调用与内容修正阶段旨在 LLMs 完成 SEMs 的自主选择，为每个子模块获取丰富且高质量的分析数据，确保报告的准确性和深度。具体来说，包括以下几个步骤：

(1) 工具调用

在工具调用步骤中，LLMs 需要根据当前子模块的任务要求 (S_{ij}, D_{ij})，以及可用的候选 SEMs 集合 E_{ij}^{match}，进行精确的调用规划。SEMs 被视为 LLMs 可调度的外部工具，LLMs 的任务是从 E_{ij}^{match} 中依次选择最合适的专家模型 e_k 用于数据获取或处理。

算法 6 大纲规划与可行性评估算法

输入: 题目模板集合 $T=\{T_1,T_2,\ldots,T_n\}$，专家集合 $E=\{E_1,E_2,\ldots,E_m\}$，大模型 LLM
输出: 最终子模块集合 $S^{\text{final}}=\{S_i^{\text{final}}\}$，对应匹配专家集合 $C=\{E_{ij}^{\text{match}}\}$
for 每个题目模板 $T_i \in T$ **do**
    $S_i = M.generate(T_i, E)$ ▷ 大纲规划
    初始化 $S_i^{\text{final}} = \emptyset$
    for 每个 $(S_{ij}, D_{ij}) \in S_i$ **do**
        $(\text{flag}_{ij}, E_{ij}^{\text{match}}) = LLM.eval(S_{ij}, D_{ij}, E)$ ▷ 可行性评估
        if $\text{flag}_{ij} = 1$ **then**
            将 (S_{ij}, D_{ij}) 加入 S_i^{final}
            保存 E_{ij}^{match} 到 C
        end if
    end for
end for

每轮调用可形式化表示为：

$$e_k = \text{LLM}_{\text{select}}(S_{ij}, D_{ij}, E_{ij}^{\text{remain}}) \tag{5-5}$$

其中，$E_{ij}^{\text{remain}} \subseteq E_{ij}^{\text{match}}$ 表示当前尚未调用的 SEMs 集合；一旦 e_k 被选中，则将其从集合中移除：

$$E_{ij}^{\text{remain}} \leftarrow E_{ij}^{\text{remain}} \setminus e_k \tag{5-6}$$

该过程持续进行，直到 $E_{ij}^{\text{remain}} = \emptyset$ 或 LLMs 返回空选择，即 $e_k = \text{None}$。通过该动态的专家调度机制，LLMs 能够充分利用每个 SEM 的专长，确保信息收集的全面性与高效性。之后 LLMs 对 E_{ij}^{call} 中的每个 SEM 发起调用请求，获取分析结果 r_{ij}：

$$r_{ij}^{(k)} = \begin{cases} E_{ij}^{(0)}(D_{ij}) & \text{若} k=1, \\ r_{ij}^{(k-1)} + E_{ij}^{(k)}(D_{ij}) & \text{若} k>1, \end{cases} \quad \text{其中} E_{ij}^{(k)} \in E_{ij}^{\text{match}}. \tag{5-7}$$

k 为调用的 SEMs 的数量，令 $E_{ij}^{\text{call}} = \{e_1, e_2, \ldots, e_K\}$，则 $E_{ij}^{(k)}$ 表示为完成任务 S_{ij} 而被选中 SEMs 集合 E_{ij}^{match} 中的第 k 个 SEM，最终得到的 $r_{ij}^{(k)} = r_{ij}$ 将作为子模块 S_{ij} 的原始数据支撑。

在工具选择机制中，本研究并未直接基于全部候选 SEMs 进行选择，而是采用“先召回、再排序”的两阶段筛选策略。原因在于，候选 SEMs 集合通常规模较大，且模型之间相似度较高，若直接从中进行选择，易导致 LLMs 在调用决策上的不稳定性，进而可能引入与当前任务紧密度较低的 SEMs，影响后续报告生成的

算法 7 工具调用与内容修正算法

输入: 子模块任务 S_{ij}，子模块数据 D_{ij}，候选专家集合 E_{ij}^{match}，大模型 LLM
输出: 优化后的分析数据 $\hat{r}_{ij}$
$E_{ij}^{\text{remain}} \leftarrow E_{ij}^{\text{match}}$ ▷ 初始化剩余专家集合
$r_{ij} \leftarrow$ "" ▷ 初始化原始分析结果
while $E_{ij}^{\text{remain}} \neq \emptyset$ **do**
    $e_k \leftarrow \text{LLM}_{\text{select}}(S_{ij}, D_{ij}, E_{ij}^{\text{remain}})$
    if $e_k = \text{None}$ **then**
        break
    end if
    $E_{ij}^{\text{remain}} \leftarrow E_{ij}^{\text{remain}} \setminus \{e_k\}$
    $r_k \leftarrow e_k(D_{ij})$ ▷ 调用处理
    $r_{ij} \leftarrow r_{ij} + r_k$ ▷ 累加分析结果
end while
$\hat{r}_{ij} \leftarrow \text{LLM}_{\text{revise}}(r_{ij})$ ▷ 内容修正

准确性与一致性。通过先行召回相关性较高的子集，再结合具体上下文进行排序与最终调用，有效提升了 SEMs 选择的相关性与稳定性，从而为报告生成提供更可靠的支撑。通过有效的工具选择，使得每个 SEM 都能在其所擅长的领域内发挥最大的效能，确保信息收集的全面性。

(2) 内容修正

在完成 SEMs 调用后，LLMs 并不会直接基于原始输出生成报告内容，而是对收集到的信息 r_{ij} 进行进一步的过滤与提炼。该过程旨在从冗长的文本或分析结果中识别出关键、高权重的信息片段，去除冗余与低质量内容，从而降低错误或无关信息对最终内容的干扰。

本研究采用直接提示（表6-6）的方式，引导 LLMs 对 SEMs 的输出进行审查、压缩和优化。该步骤可形式化为：

$$\hat{r}_{ij} = \text{LLM}_{\text{revise}}(r_{ij}) \tag{5-8}$$

其中，$\hat{r}_{ij}$ 表示子模块 S_{ij} 最终生成的内容段落，具有语言连贯、信息准确、内容紧凑等特性，有助于减少信息的冗余，并确保分析数据的高度关联与准确。

5.2.3 报告生成与优化

报告生成与优化阶段旨在通过模拟真实人类书写报告时的交互过程，将修正后的分析数据转化为结构化报告文本。该阶段的目标是确保报告表达清晰、逻辑

严谨、内容丰富，并具备较强的实用性与可读性。为此，本章研究工作引入角色协同机制，让 LLMs 分别模拟专家审核员（Reviewer）与报告撰写员（Writer）的角色，从而在交互过程中不断优化报告质量。

(1) 初稿撰写

在该阶段，审核员为每个子模块 S_{ij} 提供其标题 Title_{ij} 以及修正后的数据内容 $\hat{r}_{ij}$，作为撰写员生成初稿的输入：

$$\text{Draft}_{ij}^{(0)} = \text{LLM}_{\text{writer}}(\text{Title}_{ij}, \hat{r}_{ij}) \tag{5-9}$$

在这里，仅指定子模块的标题作为二级标题，而子模块内的报告格式则由撰写员，即由 LLMs 自主决定。这种自主性赋予了 LLMs 灵活性，使其能够根据每个子模块的具体内容来设计最合适的报告结构，确保报告内容的针对性和实用性。

(2) 审核与优化

正如一篇优秀的报告往往需要经过审核和反复修改，自动报告生成同样需要经历一个细致的迭代过程，以确保最终产出的报告达到高质量标准。在生成初稿之后，为了减少 LLMs 生成内容的不稳定性，将对报告内容进行审核与优化。

具体来说，初稿生成后，审核员对内容质量进行评估，并提出具体的修改建议 $\Delta_{ij}^{(t)}$，这些建议涵盖结构、内容准确性、语言表达、分析深度等方面。该过程形式化如下：

$$\Delta_{ij}^{(t)} = \text{LLM}_{\text{reviewer}}(\text{Draft}_{ij}^{(t)}) \tag{5-10}$$

撰写员接收建议后，基于当前版本 $\text{Draft}_{ij}^{(t)}$ 进行优化，生成下一版本：

$$\text{Draft}_{ij}^{(t+1)} = \text{LLM}_{\text{writer}}(\text{Draft}_{ij}^{(t)}, \Delta_{ij}^{(t)}) \tag{5-11}$$

该过程持续迭代，直到满足终止条件（如内容质量达标或最大轮数），最终得到高质量的子模块报告文本：

$$\text{Final}_{ij} = \text{Draft}_{ij}^{(T)} \tag{5-12}$$

所有子模块最终内容 Final_{ij} 将被汇总组成完整的结构化分析报告：

$$R_{\text{final}} = \text{Final}_{00} \parallel \text{Final}_{01} \parallel \cdots \parallel \text{Final}_{n|S_n^{\text{final}}|} \tag{5-13}$$

这里，$\parallel$ 表示字符的拼接。

通过模拟报告撰写与审核过程，不仅提升了报告内容的准确性和可读性，也增强了生成过程的稳定性与可控性。全部的提示见表6-7。

算法 8 结构化报告生成与优化算法

输入: 所有子模块标题 Title_{ij}，修正后的数据内容 $\hat{r}_{ij}$，审核员 $\text{LLM}_{\text{reviewer}}$，撰写员 $\text{LLM}_{\text{writer}}$，最大迭代轮数 T
输出: 最终结构化报告 R_{final}
$\text{R}_{\text{final}} \leftarrow$ ""
for 每个模板 i 的每个子模块 j **do**
  $\text{Draft}_{ij}^{(0)} \leftarrow \text{LLM}_{\text{writer}}(\text{Title}_{ij}, \hat{r}_{ij})$ ▷ 初稿撰写
  for $t = 0$ to $T-1$ **do**
    $\Delta_{ij}^{(t)} \leftarrow \text{LLM}_{\text{reviewer}}(\text{Draft}_{ij}^{(t)})$ ▷ 审核反馈
    if 反馈 $\Delta_{ij}^{(t)}$ 满足终止条件 **then**
      break
    end if
    $\text{Draft}_{ij}^{(t+1)} \leftarrow \text{LLM}_{\text{writer}}(\text{Draft}_{ij}^{(t)}, \Delta_{ij}^{(t)})$ ▷ 基于反馈优化
  end for
  $\text{Final}_{ij} \leftarrow \text{Draft}_{ij}^{(t)}$
  $\text{R}_{\text{final}} \leftarrow \text{R}_{\text{final}} \parallel \text{Final}_{ij}$
end for

5.3 实验与分析

5.3.1 实验设置

本节以专利技术报告的生成为例，具体阐述相关实现细节。

(1) 数据

由于资源的限制，本章研究工作专注于生成 2023 年的专利技术分析报告，从特定的在线资源[1]中收集了大量 2023 年专利数据，包括专利标题、摘要、发明人、申请公司、IPC 分类号等信息，并存储在数据库中，以便报告生成使用。为了及时捕捉技术信息，本章研究工作以双月作为一个生成周期。在这个周期内，报告覆盖了从宏观到微观、从广泛到具体的不同粒度主题，以确保进行全面而深入的实验。最终，共计生成 23 篇报告，138 个子模块。

(2) 小型专家模型

本章研究工作设计了报告生成中常见的 SEMs，包括：标签抽取 SEMs、聚类 SEMs、质量评分 SEMs、统计 SEMs。其中，标签抽取 SEMs 负责从专利文档中提取公司、专家和行业的相关技术标签，快速识别专利的技术领域和相关性；聚类 SEMs 将相似的专利归类到相应的子类别中，有助于在报告中展示专利的分布

[1] https://analysis.iprdb.com/, https://www.uyanip.com/

情况；质量评分 SEMs 通过评估专利的权利要求数量、法律状态等因素，使用岭回归模型对专利进行评分；统计 SEMs 对专利数据进行量化分析，包括计算专利的 IPC 分布、专利类型比例等，为报告提供宏观视角。

(3) 大模型

LLMs 在自动报告生成中扮演着重要的角色，负责规划和处理从 SEMs 中获得的数据，以及最终生成连贯的报告文本。考虑到实际应用的可行性和成本效益，本章研究工作选择智谱 AI[2]作为生成报告的 LLMs。智谱 AI 基于 GLM 构建，通过大量的中文数据训练和对齐，是一个在中文自然语言处理领域表现出色的生成式语言模型。

(4) 对比基线

AutoRG 旨在成为一个通用的报告生成框架，不局限于特定类型的报告，其生成的报告目录由 LLMs 动态决定，这导致在报告结构上与现有相关工作不同。同时，现有的报告生成方法通常专注于特定类型的报告，生成流程和结构目录高度定制化，使得跨方法的直接对比变得复杂。因此，本章研究工作从现有工作的工作流，即半模板式 LLMs 报告生成方法的角度进行对比。为确保比较的一致性，本章研究工作以大纲生成阶段输出的目录作为报告的标题。在半模板式 LLMs 生成方法中，报告内容的生成使用人工判定的方式，选择最适合当前子模块的一个 SEM，并将输出直接交给 LLMs 进行报告生成。

5.3.2 评估方法

考虑到一篇完整的报告内容篇幅可能过长，导致评估过程变得复杂和耗时，本章研究工作选择了以模块为单位进行评估。具体来说，根据二级标题将报告划分为多个独立的部分，每个部分都对应一个特定的报告内容。这样的划分可以使得评估过程更加高效。

(1) 主观评估

主观评价指标体系的构建参考了以往报告生成质量评估[78] 的工作，包括以下四个维度：事实性、一致性、充分性和整体质量。事实性指标旨在评估报告内容的准确性和真实性；一致性指标关注报告内部逻辑的连贯性以及与既定主题的相关性；充分性指标则评估报告内容是否全面，是否存在信息的遗漏或冗余；整体质量指标综合考量报告是否为读者提供了新的见解和知识。围绕上述评价体系，本节分别用模型评估和人为评估的方式从主观层面对报告生成的质量进行评价。

模型评估为每个评价指标设定独立的投票标准，以确保从不同维度对报告的

[2]https://www.zhipuai.cn/

质量进行细致的考量。本章研究工作使用 Sparkv3.5[3]和 Kimi[4]作为评估模型，通过两两比较的方式让评估模型进行投票。为了避免评估过程中位置造成的影响，本章研究工作随机安排了待评估段落的顺序。以半模板式报告作为基准，通过计算胜率来评估不同方法相对于半模板式报告的表现。胜率的计算公式如5-14，k 为报告数量,n_i 为第 i 篇专利比较的次数，也是子模块的数量，a_i 为被选择的数量。考虑到完整报告长度与领域性可能导致评估困难，因此本章研究工作采取了以模块为单位的整体评价的方式进行人为评估。每个评审员将随机负责 3 篇报告的对比，根据评分标准进行综合投票。

$$win = \sum_{i=1}^{k} \frac{a_i}{n_i} \tag{5-14}$$

(2) 客观评估

为降低主观性对评估结果的影响，本章研究工作进一步设置了不同的客观评估指标。这些指标涵盖了词汇多样性、内容联系度、领域独特性和句法复杂性四个方面，旨在从更丰富的维度对生成报告进行客观对照。词汇多样性的评估采用了类符-形符比（Type-Token Ratio, TTR）作为衡量指标。TTR 通过计算文本中不同词汇类型（Type）与总词数（Token）的比率，来量化文本的词汇丰富度。这一指标能够有效反映报告中词汇的使用广度和多样性。

内容联系度和领域独特性的评估则采用了 self-BLEU 方法。其中，内容联系度旨在衡量一篇报告内部各个模块之间的联系是否紧密，通过比较报告内部模块与其他模块之间的相似性，来衡量报告整体的联系度。领域独特性旨在评估不同领域之间报告与报告的相似性，如果对于同一个模块，不同领域报告之间的内容过于相似，这样的报告在领域深度上有较大的提升空间。

句法复杂性的评估基于句子嵌套深度的计算。通过测量句子中嵌套句法结构的最大深度，可以评估句子的句法复杂度，进而反映句子结构的复杂性，以及生成文本的句法多样性。

$$TTR = \frac{\#\ \text{Type}}{\#\ \text{Token}} \tag{5-15}$$

$$self-BLEU = exp(-\sum_{i=1}^{n} w_i \cdot \log \frac{C(i)}{Ni}) \tag{5-16}$$

[3]https://xinghuo.xfyun.cn/

[4]https://kimi.moonshot.cn/

5.3.3 实验结果分析

(1) 主观评估

1) AutoRG 生成的报告在内容上更加充分、全面。如表5-1所示，模型评估结果揭示了 AutoRG 与半模板式报告方法相比的胜率，其中高亮显示的部分表示 AutoRG 的平均胜率超过了半模板式 LLMs 报告生成方法。特别是在报告的充分性这一关键指标上，AutoRG 展现出了显著的优越性。这归功于 AutoRG 利用 LLMs 的自主规划能力，能够自主选择并整合多种 SEMs，生成内容更为全面和详尽的报告，证实了 AutoRG 框架在增强报告内容丰富度方面的显著效果。附录6.2也给出了半模板与 AutoRG 在不同领域下的报告样例。以生成“龙头企业专利质量分析”的图6-1为例，半模板式方法生成的报告只是单一的不同公司的专利介绍的罗列，而 AutoRG 通过 LLM 思考，调用不同的 SEMs，从专利 IPC、专利类型不同方面，结合各公司的优质专利技术发布动态情况，对企业专利进行分析，有助于为读者提供更深入的思考。

2) AutoRG 生成的报告在质量上更稳定，效果更好。评估结果显示，AutoRG 在事实性、一致性以及整体质量方面均超越了传统的半模板式报告方法，在报告质量上有着一定的优势。此外，通过箱型图（见图5-3）对生成报告的胜率分布进行可视化分析，可以观察到 AutoRG 在所有评估指标的胜率分布上均显著高于半模板式报告，并且在一致性、充分性和整体质量中均呈左偏分布，这表明在多数情况下，AutoRG 能够产生质量较高的报告。这一统计特性凸显了 AutoRG 在维持报告质量方面的出色能力。

表 5-1 不同报告生成方法下胜率的对比

Table 5-1 Comparative Analysis of Success Rates Across Automated Report Generation Methodologies

	事实性			一致性			充分性			整体质量		
	Spark	Kimi	avg	Spark	Kimi	avg	Spark	Kimi	avg	Spark	Kimi	avg
AutoRG	**61.59**	**61.59**	**61.59**	**50.72**	60.14	**55.43**	**64.49**	64.49	**64.49**	**60.14**	**63.77**	**61.96**
w/o 工具选择	58.54	56.91	57.72	39.84	58.54	49.19	57.72	63.41	60.57	56.10	59.35	57.72
w/o 工具调用规划	60.98	54.47	57.72	47.15	52.85	50.00	64.23	58.54	61.38	57.72	57.72	57.72
w/o 自主信息修正	60.98	60.16	60.57	49.59	60.16	54.88	59.35	**65.04**	62.20	55.28	62.60	58.94
w/o 内容审核迭代	53.03	60.61	56.82	43.94	**60.61**	52.27	53.79	62.12	57.95	49.24	62.12	55.68

同时，本章研究工作将人为评估结果通过网页形式进行了直观展示[5]。具体来说，图5-4展示了评审员对投票结果的统计分析。在这项评估中，每位评审员均随机针对评估规则对一篇报告的不同模块分别进行了投票。图中可以清晰地观察到 AutoRG 在人为评估中获得了压倒性的优势，这一结果进一步证实了 AutoRG 的有

[5]http://8.130.143.149:5000/

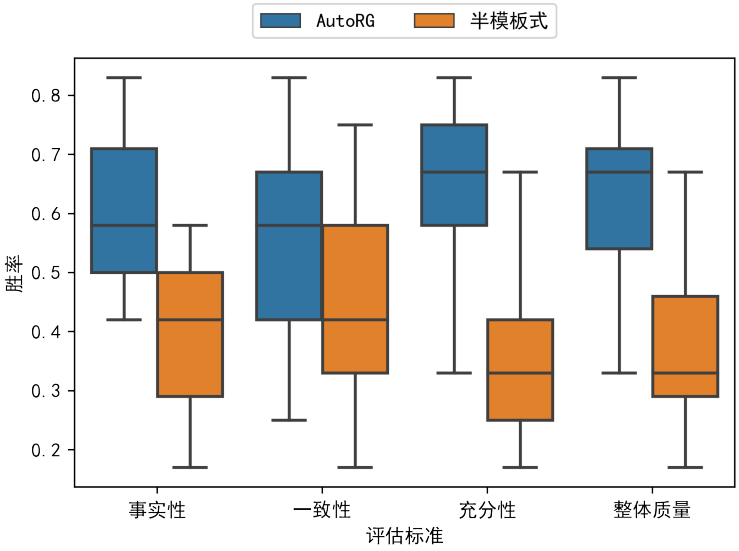


图 5-3 不同维度下的箱型图对比

Fig. 5-3 Comparison of boxplots in different dimensions

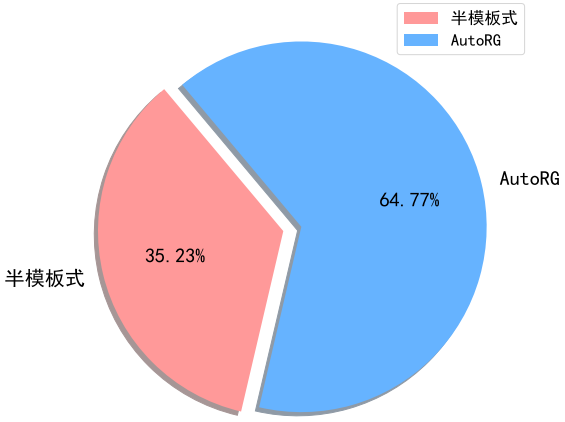


图 5-4 人为评估投票结果

Fig. 5-4 Human-Evaluated Voting Outcomes

效性和优越性。

(2) 客观评估

1) AutoRG 生成的报告在词汇运用上更加多样。如表5-3所示，AutoRG 在 TTR 指标上明显超越了传统的半模板式方法。这一结果揭示了 AutoRG 在词汇选择上的丰富性，更有效地传达了信息。这种词汇多样性的实现主要归功于在工具调用、内容修正与迭代阶段的设计。AutoRG 不仅能自主选择与任务最匹配的多样化 SEMs，还能通过信息修正精准提炼关键信息，运用更为丰富和多变的词汇进行表达。这一过程增强了报告的准确性，也提升了其可理解性。而迭代过程的优化则进一步增强了报告的词汇丰富度与表达能力。

2) AutoRG 生成的报告在模块间建立了更紧密的联系。实验数据显示，AutoRG 生成的报告在报告内 self-BLEU 指标上表现出更高的相似度。这得益于规划与工

表 5-2 消融实验下胜率的变化

Table 5-2 Performance Variation Analysis in Ablation Studies

	事实性	一致性	充分性	整体质量
AutoRG	-	-	-	-
w/o 工具选择	↓3.87	↓**6.24**	↓3.92	↓4.24
w/o 工具调用规划	↓3.87	↓5.43	↓3.11	↓4.24
w/o 自主信息修正	↓1.02	↓0.55	↓2.29	↓3.02
w/o 内容审核迭代	↓**4.77**	↓3.16	↓**6.54**	↓**6.28**

表 5-3 客观指标下不同报告生成方法的对比

Table 5-3 Comparative Evaluation of Report Generation Methodologies Using Objective Metrics

	TTR	报告内 self-BLEU	报告间 self-BLEU	句法深度
半模板式	0.4805	0.0216	0.0889	4.7793
AutoRG	0.6066	0.0229	0.0830	4.7889

具调用，使得相同的 SEMs 信息在多个模块之间流畅传递。这意味着报告中的各个模块不再是孤立的信息点，而是相互关联、相互支撑的整体。这样的结构使得读者在阅读报告时能够更容易地跟踪信息的脉络，理解各个部分之间的逻辑关系，从而更全面地把握报告的核心内容。

3) AutoRG 生成的报告在不同领域间展现出更高的适应性与差异性。通过对比报告间 self-BLEU 指标，可以发现，AutoRG 在领域多样性方面具有一定的优势。报告之间更高的差异性表明能够根据不同领域的需求，提供更为精准和独特的分析报告。这种差异性依然得益于信息修正模块与报告迭代过程中的细化调整。通过模拟人类专家的审核与优化过程，自动式生成能够对报告进行定制化调整，提高了报告的专业程度，也确保了其在不同领域间的差异性和适用性。

4) AutoRG 生成的报告更具专业性。表5-3中，通过测量句法深度，可以发现 AutoRG 生成的报告在句法结构的复杂性上表现出细微的优势。这种相对复杂的句子表达说明 AutoRG 在分析内容上相比半模板式在内容上更为透彻和深入，使得生成的报告在专业性上可能更胜一筹。

(3) 消融实验

为验证方法中不同模块的贡献，本研究工作设置了消融实验。分别通过与不进行可行性评估中的工具选择、不进行工具调用规划、不做内容修正以及不进行审核优化下生成的报告进行对比，直观分析这四个部分对最终报告质量的影响。结果如表5-2所示，有如下几点结论：

1) 多轮工具选择对提升报告一致性有积极作用。在可行性评估阶段，工具选择和工具调用规划对一致性的影响尤为关键。直接使用所有 SEMs 作为候选工具集会导致一致性下降 6.24%，而将初轮工具选择的结果直接作为候选工具集则导致 5.43% 的下降。这归结为对庞大工具集的单轮选择虽然一定程度确保了信息的

多样性，但也降低了信息的联系度，削弱了子模块与分析内容之间的相关性。而本章研究工作采用先粗后精的方式，通过初步筛选和基于详细描述的进一步筛选，提升了报告的一致性和连贯性。

2) 输入数据的质量影响报告的整体质量。自主信息修正模块的作用是对 SEMs 的原始输出做信息的过滤与整合。从表5-2的结果可以发现，使用初始 SEMs 的输出作为输入数据会导致报告在分析层面的质量下降，这一点在报告整体质量的显著降低中得到了体现。在这一过程中，过多的非相关信息会干扰分析过程，造成输入的冗长和不集中。自主信息修正模块可以进行有效的数据净化，提高报告的分析深度。

3) 迭代对报告内容的优化至关重要。与其他模块相比，当取消内容审核迭代阶段，生成报告的四个指标均有更显著的下降，这揭示了迭代对报告内容优化的重要性。迭代提供了精炼报告内容的机会，允许对初版中可能被忽视的细节和未充分探讨的分析进行深入审视。这种修正和完善的过程，确保了最终报告能够全面且准确地反映分析结果，提高了报告的质量。

5.4 本章小结

本章以大小模型协同框架的应用为出发点，介绍了一个大小模型协同的自动报告生成框架 AutoRG。AutoRG 通过有效利用 LLMs 的工具理解与自主规划能力，自动化实现更灵活和全面的报告。同时，通过引入信息修正机制与报告迭代机制，降低对 SEMs 输出和 LLMs 生成随机性的依赖，提高了报告的质量与稳定性。本章研究工作以专利技术报告生成为例，通过模型主观评估和客观评估结合的方式，从多个维度证明，AutoRG 生成的报告在在事实性、一致性、充分性和整体质量上有着更优的效果。

6 总结与展望

6.1 总结

随着 LLMs 能力的快速提升，其在自然语言推理与生成任务中的表现不断逼近甚至超过人类水平。然而，在特定垂直领域应用中，LLMs 仍面临专业性不足、更新滞后、知识构建困难等问题。SEMs 则具备较强的领域能力、灵活的部署等优势，可以很好的适配 LLMs 的推理框架。因此，如何实现 LLMs 与 SEMs 的协同推理，成为提升领域性能的关键问题。

本文围绕“大小模型协同推理”的研究主题，从理论与实践两个层面出发。在理论层面，分别在有标签与无标签场景中优化协同推理方法，提升大小模型协同在领域任务中的性能；在实践层面，本文设计了一个完整的协同框架，应用于自动报告生成任务，验证多 SEMs 场景下的协同效果，为缓解 LLMs 领域知识不足的问题提供了行之有效的解决方案。具体工作总结如下：

(1) 在有标签场景下，针对 SEMs 置信度有偏差、预测错误时的优化策略问题，提出基于概率校准与错误驱动的协同推理方法。通过对 SEMs 输出进行 Beta 概率校准，提高了置信度的准确性；并通过对错误样本进行挖掘，构建错误驱动的推理原则，帮助 LLMs 在 SEMs 发生错误时，依然可以正确推理。实验结果显示，该方法在新闻分类、生物医学问答等任务中，不仅提高了分类准确性和 F1 值，还显著改善了推理的鲁棒性。

(2) 在无标签场景下，为解决 SEMs 的偏好知识缺乏生成验证机制的问题，构建了基于局部假设辩论与全局聚类的双粒度知识验证体系。在局部层面，通过增加候选答案和大模型决策机制实现共识局部知识生成；在全局层面，借助聚类对领域共性知识进行提炼，进行补充和校验。实验结果表明，该策略在不同任务和数据条件下均能有效增强 LLMs 的稳定性和泛化能力，特别在低资源环境下体现出了较强的鲁棒性，有效缓解了错误信息的扩散。

(3) 面向实际应用，本文进一步构建了一个多 SEMs 协同的自动报告生成框架 AutoRG，实现从大纲生成到报告成稿的完整流程。该框架通过先召回后排序的方式协调多个 SEMs 的调用，并通过引入信息修正与报告迭代机制，显著缓解了 SEMs 误差和 LLMs 输出波动对报告质量的影响，最终实现了自动化程度高、生成质量稳定的报告生成效果。

6.2 展望

尽管本文工作在协同推理与自动报告生成等问题上取得了积极成效，但在实际应用和理论完善上仍存在若干不足，有待后续深入研究与改进。

首先，当前提出的融合推理框架在准确性方面虽实现了明显提升，但“以专助通”的协同目标尚未完全达成，尤其在某些任务和某些 LLMs 上尚未实现对 SEMs 的全方位性能超越，仍需在信息交互、协同策略设计等方面持续改进，以进一步挖掘 LLMs 的通用能力与 SEMs 的专业性之间的协同潜力。

其次，当前的融合推理框架引入 SEMs 在提升推理效果的同时，也引入了额外的计算开销。例如，在无标签场景中，局部辩论与全局聚类等模块的推理较复杂，尤其在大规模数据处理时，容易成为系统运行效率的瓶颈。未来可能设计自适应的协同机制，使得模型能够根据当前任务复杂度和实时计算资源动态调整协同力度。对于简单任务，优先调动效率更高的模型完成推理；而在面对复杂或低置信度样本时，再引入更多专家模块进行补充。这样的动态策略既能保证整体性能，又能有效降低资源浪费。

最后，大小模型协同的核心在于实现知识的有效传递与能力互补，充分发挥 LLMs 的通用能力与 SEMs 的精细调优能力。但实际场景中，并非所有领域都存在合适的 SEMs。在跨任务和跨领域应用场景下，如何保证两类模型间信息传递的准确性和一致性，仍然面临挑战和发展方向。对于跨任务的信息传递，未来可以探索构建任务无关的中间表示或知识图谱，将通用知识与任务专属信息进行有效融合，使得大小模型在面对多样化任务时能够更好地对齐信息；对于跨领域的信息传递，未来工作可从领域自适应技术出发，借助对比学习、领域自编码器等方法，将领域无关的信息提取出来，而将领域特定信息进行单独建模，达到知识分离与融合的目的，使得大小模型的知识表征能够在不同领域间进行映射和调整，从而实现更好的对齐。

参考文献

[1] Louis A, van Dijck G, Spanakis G. Interpretable Long-form Legal Question Answering with Retrieval-augmented Large Language Models [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 22266–22275.

[2] 沈晨晨, 岳圣斌, 刘书隽, et al. 面向法律领域的大模型微调与应用. [J]. Big Data Research, 2024, 10 (5): 11–27.

[3] Singhal K, Tu T, Gottweis J, et al. Toward Expert-level Medical Question Answering with Large Language Models [J]. Nature Medicine, 2025: 1–8.

[4] 陈晓红, 刘浏, 袁依格, et al. 医疗大模型技术及应用发展研究 [J]. 中国工程科学, 2024, 26 (6): 77–88.

[5] 润生陈. 医疗大数据结合大语言模型的应用展望 [J]. Journal of Sichuan University (Medical Sciences), 2023, 54 (5): 855.

[6] Aguda T D, Siddagangappa S, Kochkina E, et al. Large Language Models as Financial Data Annotators: A Study on Effectiveness and Efficiency [C] // Proceedings of the 2024 Joint International Conference on Computational Linguistics Language Resources and Evaluation (LREC-COLING 2024), 2024: 10124–10145.

[7] 程大伟, 贾仁军, 李江彤, et al. 知识增强的中文金融大模型研究. [J]. Big Data Research (2096-0271), 2025, 11 (2): 5–18.

[8] 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.

[9] 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.

[10] Liu A, Yang Z, Zhang Z, et al. PANDA: Preference Adaptation for Enhancing Domain-Specific Abilities of LLMs [C] // Findings of the Association for Computational Linguistics: ACL, 2024: 10960–10977.

[11] Li H, Ai Q, Chen J, et al. Blade: Enhancing Black-box Large Language Models with Small Domain-specific Models [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2025: 24422–24430.

[12] Topsakal O, Akinci T C. Creating Large Language Model Applications Utilizing Langchain: A Primer on Developing Llm Apps Fast [C] // International Conference on Applied Engineering and Natural Sciences, 2023: 1050–1056.

[13] Schick T, Dwivedi-Yu J, Dessi R, et al. Toolformer: Language Models Can Teach Themselves to Use Tools [J], 2023, 36: 68539–68551.

[14] Ruan J, Chen Y, Zhang B, et al. TPTU: Large Language Model-based AI Agents for Task Planning and Tool Usage [J]. arXiv preprint arXiv:2308.03427, 2023.

[15] Ling C, Zhao X, Lu J, et al. Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey [J]. arXiv preprint arXiv:2305.18703, 2023.

[16] Radford A, Wu J, Child R, et al. Language Models are Unsupervised Multitask Learners [J]. OpenAI blog, 2019, 1 (8): 9.

[17] Agrawal M, Hegselmann S, Lang H, et al. Large Language Models are Few-shot Clinical Information Extractors [C] // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 1998–2022.

[18] Lee D-H, Kadakia A, Tan K, et al. Good Examples Make A Faster Learner: Simple Demonstration-based Learning for Low-resource NER [C] // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 2687–2700.

[19] Ferber D, Wölflein G, Wiest I C, et al. In-context Learning Enables Multimodal Large Language Models to Classify Cancer Pathology Images [J]. Nature Communications, 2024, 15 (1): 10104.

[20] Mo Y, Liu J, Yang J, et al. C-ICL: Contrastive In-context Learning for Information Extraction [C] // Findings of the Association for Computational Linguistics: EMNLP, 2024: 10099–10114.

[21] Zhang M, Gautam V, Wang M, et al. The Impact of Demonstrations on Multilingual In-Context Learning: A Multidimensional Analysis [C] // Findings of the Association for Computational Linguistics: ACL, 2024: 7342–7371.

[22] 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.

[23] Xiang Y, Yan H, Gui L, et al. Addressing Order Sensitivity of In-Context Demonstration Examples in Causal Language Models [C] // Findings of the Association for Computational Linguistics: ACL, 2024: 6467–6481.

[24] Liu H, Yin H, Luo Z, et al. Integrating Chemistry Knowledge in Large Language Models via Prompt Engineering [J]. Synthetic and Systems Biotechnology, 2025, 10 (1): 23–38.

[25] 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.

[26] Yao Y, Li Z, Zhao H. Beyond Chain-of-thought, Effective Graph-of-thought Reasoning in Language Models [J]. arXiv preprint arXiv:2305.16582, 2023.

[27] Zhou Y, Geng X, Shen T, et al. Thread of Thought Unraveling Chaotic Contexts [J]. arXiv preprint arXiv:2311.08734, 2023.

[28] Kandpal N, Deng H, Roberts A, et al. Large Language Models Struggle to Learn Long-tail Knowledge [C] // International Conference on Machine Learning, 2023: 15696–15707.

[29] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented Generation for Knowledge-intensive Nlp Tasks [J]. Advances in Neural Information Processing Systems, 2020, 33: 9459–9474.

[30] Li D, Yan J, Zhang T, et al. On the Role of Long-tail Knowledge in Retrieval Augmented Large Language Models [C] // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 120–126.

[31] Xia Y, Wu J, Kim S, et al. Knowledge-aware Query Expansion with Large Language Models for Textual and Relational Retrieval [J]. arXiv preprint arXiv:2410.13765, 2024.

[32] Asai A, Wu Z, Wang Y, et al. Self-rag: Learning to Retrieve, Generate, and Critique through Self-reflection [C] // The Twelfth International Conference on Learning Representations, 2023.

[33] Ram O, Levine Y, Dalmedigos I, et al. In-context Retrieval-augmented Language Models [J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1316–1331.

[34] Wang F, Wan X, Sun R, et al. Astute Rag: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models [J]. arXiv preprint arXiv:2410.07176, 2024.

[35] Xu R, Qi Z, Guo Z, et al. Knowledge Conflicts for LLMs: A Survey [C] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 8541–8565.

[36] Wu K, Wu E, Zou J Y. Clasheval: Quantifying the Tug-of-war between an Llm' s Internal Prior and External Evidence [J]. Advances in Neural Information Processing Systems, 2024, 37: 33402–33422.

[37] Xi Z, Chen W, Guo X, et al. The Rise and Potential of Large Language Model based Agents: A Survey [J]. Science China Information Sciences, 2025, 68 (2): 121101.

[38] Cui J, Ning M, Li Z, et al. Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model. 2023.

[39] Shen Y, Song K, Tan X, et al. Hugginggpt: Solving Ai Tasks with Chatgpt and Its Friends in Hugging Face [J]. Advances in Neural Information Processing Systems, 2024, 36.

[40] 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.

[41] Wang W, Dong L, Cheng H, et al. Augmenting Language Models with Long-term Memory [J]. Advances in Neural Information Processing Systems, 2023, 36: 74530–74543.

[42] Hsieh C-Y, Li C-L, Yeh C-k, et al. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes [C] // Findings of the Association for Computational Linguistics: ACL, 2023: 8003–8017.

[43] Li S, Chen J, Chen Z, et al. Explanations from Large Language Models Make Small Reasoners Better [C] // 2nd Workshop on Sustainable AI.

[44] Huang Y, Chen Y, Yu Z, et al. In-context Learning Distillation: Transferring Few-shot Learning Ability of Pre-trained Language Models [J]. arXiv preprint arXiv:2212.10670, 2022.

[45] Qin C, Xia W, Jiao F, et al. Improving In-context Learning via Bidirectional Alignment [J]. arXiv preprint arXiv:2312.17055, 2023.

[46] Agarwal R, Vieillard N, Stanczyk P, et al. GKD: Generalized Knowledge Distillation for Auto-regressive Sequence Models. 06 2023.

[47] Kang M, Lee S, Baek J, et al. Knowledge-augmented Reasoning Distillation for Small Language Models in Knowledge-intensive Tasks [J]. Advances in Neural Information Processing Systems, 2023, 36: 48573–48602.

[48] Lin B Y, Fu Y, Yang K, et al. Swiftsage: A Generative Agent with Fast and Slow Thinking for Complex Interactive Tasks [J]. Advances in Neural Information Processing Systems, 2023, 36: 23813–23825.

[49] Tan J, Dou Z, Zhu Y, et al. Small Models, Big Insights: Leveraging Slim Proxy Models To Decide When and What to Retrieve for LLMs [C] // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 4420–4436.

[50] Mekala D, Nguyen A, Shang J. Smaller Language Models are Capable of Selecting Instruction-Tuning Training Data for Larger Language Models [C] // Findings of the Association for Computational Linguistics: ACL, 2024: 10456–10470.

[51] Li M, Zhang Y, He S, et al. Superfiltering: Weak-to-Strong Data Filtering for Fast Instruction-Tuning [C] // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 14255–14273.

[52] Xia M, Malladi S, Gururangan S, et al. LESS: Selecting Influential Data for Targeted Instruction Tuning [C] // Proceedings of the 41st International Conference on Machine Learning, 2024: 54104–54132.

[53] Xu C, Xu Y, Wang S, et al. Small Models are Valuable Plug-ins for Large Language Models [C] // Findings of the Association for Computational Linguistics: ACL, 2024: 283–294.

[54] Yang L, Zhang S, Yu Z, et al. Supervised Knowledge Makes Large Language Models Better In-context Learners [C] // The Twelfth International Conference on Learning Representations.

[55] Burns C, Izmailov P, Kirchner J H, et al. Weak-to-strong Generalization: Eliciting Strong Capabilities with Weak Supervision [C] // Proceedings of the 41st International Conference on Machine Learning, 2024: 4971–5012.

[56] Ji J, Chen B, Lou H, et al. Aligner: Efficient Alignment by Learning to Correct [J]. Advances in Neural Information Processing Systems, 2024, 37: 90853–90890.

[57] Liu Y, Alahi A. Co-Supervised Learning: Improving Weak-to-Strong Generalization with Hierarchical Mixture of Experts [J]. CoRR, 2024.

[58] Ormazabal A, Artetxe M, Agirre E. CombLM: Adapting Black-Box Language Models through Small Fine-Tuned Models [C] // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 2961–2974.

[59] Zhao T, Wei M, Preston J S, et al. Automatic Calibration and Error Correction for Large Language Models via Pareto Optimal Self-supervision [J]. arXiv preprint arXiv:2306.16564, 2023.

[60] Xu W, Agrawal S, Briakou E, et al. Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection [J]. Transactions of the Association for Computational Linguistics, 2023, 11: 546–564.

[61] Ji J, Chen B, Lou H, et al. Aligner: Achieving Efficient Alignment through Weak-to-strong Correction [J]. arXiv e-prints, 2024: arXiv–2402.

[62] Guo C, Pleiss G, Sun Y, et al. On Calibration of Modern Neural Networks [C] // International Conference on Machine Learning, 2017: 1321–1330.

[63] Guo Y, Yang Y. Improving Weak-to-strong Generalization with Reliability-aware Alignment [J]. arXiv preprint arXiv:2406.19032, 2024.

[64] Minsky M P S. An Introduction to Computational Geometry [J]. Cambridge Tiass., HIT, 1969, 479(480): 104.

[65] Zhang T, Madaan A, Gao L, et al. In-context Principle Learning from Mistakes [C] // Proceedings of the 41st International Conference on Machine Learning, 2024: 59520–59558.

[66] Sun H, Jiang Y, Wang B, et al. Retrieved In-Context Principles from Previous Mistakes [C] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 8155–8169.

[67] Ran X, Xi Y, Lu Y, et al. Comprehensive Survey on Hierarchical Clustering Algorithms and the Recent Developments [J]. Artificial Intelligence Review, 2023, 56 (8): 8219–8264.

[68] Huang W, Chen E, Liu Q, et al. Hierarchical Multi-label Text Classification: An Attention-based Recurrent Network Approach [C] // Proceedings of the 28th ACM international conference on information and Knowledge management, 2019: 1051–1060.

[69] yeon Sung Y, Kim S B. Topical Keyphrase Extraction with Hierarchical Semantic Networks [J]. Decision Support Systems, 2020, 128: 113163.

[70] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network [J]. stat, 2015, 1050: 9.

[71] Leike J, Krueger D, Everitt T, et al. Scalable Agent Alignment via Reward Modeling: A Research Direction [J]. arXiv preprint arXiv:1811.07871, 2018.

[72] Kenton Z, Siegel N, Kramár J, et al. On Scalable Oversight with Weak Llms Judging Strong Llms [J]. Advances in Neural Information Processing Systems, 2024, 37: 75229–75276.

[73] Burns C, Izmailov P, Kirchner J H, et al. Weak-to-strong Generalization: Eliciting Strong Capabilities with Weak Supervision [C] // Proceedings of the 41st International Conference on Machine Learning, 2024: 4971–5012.

[74] Ma Y, Cao Y, Hong Y, et al. Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples! [C] // Findings of the Association for Computational Linguistics: EMNLP, 2023: 10572–10601.

[75] Nakaura T, Yoshida N, Kobayashi N, et al. Preliminary Assessment of Automated Radiology Report Generation with Generative Pre-trained Transformers: Comparing Results to Radiologist-generated Reports [J]. Japanese Journal of Radiology, 2024, 42 (2): 190–200.

[76] Ren Y, Hu W, Wang Z, et al. A Hybrid Deep Generative Neural Model for Financial Report Generation [J]. Knowledge-Based Systems, 2021, 227: 107093.

[77] Chapman C L, Hillebrand L, Stenzel M R, et al. Towards Generating Financial Reports from Tabular Data Using Transformers [C] // International Cross-Domain Conference for Machine Learning and Knowledge Extraction, 2022: 221–232.

[78] Babour A, Khan J I. Automatic Generation of a Metric Report: A Case Study of Scientometric Analytics [J]. IEEE Access, 2021, 10: 3923–3934.

[79] Ding Y, Ma Y, Fan W, et al. Fashionregen: Llm-empowered Fashion Report Generation [C] // Companion Proceedings of the ACM Web Conference 2024, 2024: 991–994.

[80] Wang J, Zhang H, Zhang C, et al. An Effective Scheme for Generating An Overview Report over A Very Large Corpus of Documents [C] // Proceedings of the ACM Symposium on Document Engineering 2019, 2019: 1–11.

[81] Gong J, Ren W, Zhang P. An Automatic Generation Method of Sports News based on Knowledge Rules [C] // 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017: 499–502.

[82] Noh Y, Shin Y, Park J, et al. WIRE: An Automated Report Generation System Using Topical and Temporal Summarization [C] // Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020: 2169–2172.

[83] Li M, Liu R, Wang F, et al. Auxiliary Signal-guided Knowledge Encoder-decoder for Medical Report Generation [J]. World Wide Web, 2023, 26 (1): 253–270.

[84] Tsaniya H, Fatichah C, Suciati N. Automatic Radiology Report Generator Using Transformer with Contrast-based Image Enhancement [J]. IEEE Access, 2024.

[85] Li C Y, Liang X, Hu Z, et al. Knowledge-driven Encode, Retrieve, Paraphrase for Medical Image Report Generation [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 6666–6673.

[86] Li Q, Xu J, Yuan R, et al. Enhanced Knowledge Injection for Radiology Report Generation [C] // 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2023: 2053–2058.

[87] Wang Z, Ren Y, Zhang X, et al. Generating Long Financial Report Using Conditional Variational Autoencoders with Kowledge Distillation [J]. IEEE Transactions on Artificial Intelligence, 2024.

[88] Colverd G, Darm P, Silverberg L, et al. FloodBrain: Flood Disaster Reporting by web-based Retrieval Augmented Generation with An LLM [C] // Artificial Intelligence for Humanitarian Assistance and Disaster Response Workshop, 2023.

[89] Reddy R G, Fung Y R, Zeng Q, et al. Smartbook: Ai-assisted Situation Report Generation [J]. arXiv preprint arXiv:2303.14337, 2023.

[90] 曾文龙, 刘丹, 张超. 基于大模型的智能抄清: 事件要点抽取与报告生成. [J]. Cyber Security & Data Governance, 2023, 42 (12).

附录 A

基于概率校准与错误驱动的监督协同推理算法在 SST-5 中的指令

表 6-1 错误驱动的原则生成提示模版

Table 6-1 Prompt template of the Generation Based on Error-driven Principles Module

初始原则生成 Ψ_{origin}	# Instruction Provide a robust insight from mistakes to enhance understanding of categories in Ethical Classification. Input: {question} Generated Category: {generated_answer} Correct Category: {correct_answer} # Note We are not focused on this one data point, but rather on the general and typical principle. Inputs are variable, but insight is robust. # Output Format Insights: <provide one robust principle from mistakes .> # Example Insights: <POSSIBLY_NEEDS_*> suggests a level of uncertainty, whereas <NEEDS_*> appropriately conveys the necessity for careful consideration. Focus on the severity and immediacy of the issue.
原则迭代 Ψ_{update}	# Instruction Supplement or optimize the old insights from mistakes to enhance the understanding of the emotion classification. Task Input: {question} Generated Category: {generated_answer} Correct Category: {correct_answer} # Note We are not focused on this one data point, but rather on the general and typical principle. Context and question are variable, but insight is robust. Pay attention to the simplification of the content. # Output Format Insights: <provide one robust principle from mistakes .> # Old Insight Insights: {insight}

表 6-2 各方法的推理提示模版

Table 6-2 Reasoning Prompt Templates for each Method

CCDC 推理	# Task: Emotion classification and determine the emotion of the last sentence. You can use an expert model to aid your judgment, and output the correct emotion for the last sentence. The expert's confidence represents its certainty in providing the answer. {DIAG} Note: Choose from the following emotions: <VERY_NEGATIVE>, <NEGATIVE>, <NEUTRAL>, <POSITIVE>, <VERY_POSITIVE> and only output the emotion. # Example {CONTEXT} # Test {TEST}
SuperICL 推理	# Task: Emotion classification and determine the emotion of the last sentence. You can use an expert model to aid your judgment, and output the correct emotion for the last sentence. The expert's confidence represents its certainty in providing the answer. Note: Choose from the following emotions: <VERY_NEGATIVE>, <NEGATIVE>, <NEUTRAL>, <POSITIVE>, <VERY_POSITIVE> and only output the emotion. # Example {CONTEXT} # Test {TEST}
SuperContext 推理	Use the prediction from the expert model as a reference to aid your judgment. {CONTEXT} Data: {TEST} Please provide your analysis using the format below and then give your final prediction: 1. Influence Degree: On a scale of 0 to 1 (in increments of 0.1), how much did the expert's prediction influence your judgment? 2. Critical Features: Identify the specific word or phrase in the test case that played a pivotal role in your prediction of emotion classification. 3. Final Prediction: Choose from the following emotions: <VERY_NEGATIVE>, <NEGATIVE>, <NEUTRAL>, <POSITIVE>, <VERY_POSITIVE>.

表 6-2（续表）

ICL 推理	# Task: Emotion classification and determine the emotion of the last sentence. Note: Choose from the following emotions: <VERY_NEGATIVE>, <NEGATIVE>, <NEUTRAL>, <POSITIVE>, <VERY_POSITIVE> and only output the emotion. # Example {CONTEXT} # Test {TEST}

附录 B

基于多源知识验证的无监督协同推理算法在 SST-5 中的指令

表 6-3 粗细粒度知识生成的提示模版

Table 6-3 Prompt Templates for Generating Coarse and Fine-grained Knowledge

全局知识 生成 Ψ_{cluster}	There is an expert who can determine the sentiment of given text and return an integer based on {{VERY_NEGATIVE: 0, NEGATIVE: 1, NEUTRAL: 2, POSITIVE: 3, VERY_POSITIVE: 4}}. Generate a knowledge within 64 words to tell me how to classifiy the sentiment accurately based on the following predictions: {TRAIN} Start with “When the scenario is about ... , to determine the sentiment of a given text,”.
候选论点 生成 Ψ_{argue}	There is an expert who can determine the sentiment of given text and return an integer based on {{VERY_NEGATIVE: 0, NEGATIVE: 1, NEUTRAL: 2, POSITIVE: 3, VERY_POSITIVE: 4}}. Here is the expert ’s prediction: Text: {txt} Answer (0 or 1 or 2 or 3 or 4): {first}({fid}) Please generate a **brief short** knowledge within 64 words to support why the expert prefers {first}({fid}) rather than {second}({sid}). Start with “To determine the sentiment of a given text, we should pay attention to ...”.
多假设 决策 Ψ_{judge}	Text: {txt} Analyze which of the following insight is more accurate. (Insight_a): {ia} (Insight_b): {ib} Don’t be influenced by the length of insight. Don’t be fooled by false insight. If (Insight_a) is more accurate, answer ‘(Insight_a)’. If (Insight_b) is more accurate, answer ‘(Insight_b)’.

表 6-4 各方法的推理提示模版

Table 6-4 Reasoning Prompt Templates for each Method

Dual-Know 推理	Output the sentiment of the given text. Choose your answer from provided list and map your answer with following {{VERY_NEGATIVE: 0, NEGATIVE: 1, NEUTRAL: 2, POSITIVE: 3, VERY_POSITIVE: 4}} and return an integer as a result. There are some information that may be helpful: {GLOBAL} {INSIGHTS} {TEST}
Panda 推理	Output the sentiment of the given text. Choose your answer from provided list and map your answer with following {{VERY_NEGATIVE: 0, NEGATIVE: 1, NEUTRAL: 2, POSITIVE: 3, VERY_POSITIVE: 4}} and return an integer as a result. There are some information that may be helpful: {INSIGHTS} {TEST}
ICL 或 ICL+Conf 推理	Output the sentiment of the given text. Choose your answer from provided list and map your answer with following {{VERY_NEGATIVE: 0, NEGATIVE: 1, NEUTRAL: 2, POSITIVE: 3, VERY_POSITIVE: 4}} and return an integer as a result. There are some information that may be helpful: {TRAIN} {TEST}
ICL+CoT 推理	Output the sentiment of the given text. Choose your answer from provided list and map your answer with following {{VERY_NEGATIVE: 0, NEGATIVE: 1, NEUTRAL: 2, POSITIVE: 3, VERY_POSITIVE: 4}} and return an integer as a result. There are some information that may be helpful: {TRAIN} {TEST} Think about it step by step:

附录 C

一种大小模型协同的自动报告生成框架指令

表 6-5 大纲规划与评估的大模型提示

Table 6-5 Prompt of the Outline Planning and Evaluation Module

大纲规划	如果生成 “{domain} 专利的技术分析报告’’，对于其中的 “{outline}’’ 模块，可能包括哪些部分? 分别需要哪些数据? 要求的输出：1. 分点回答；2. 格式是 “1. xx 分析”
可行性评估	# 你的任务是生成一个名为 “{domain} 专利技术分析报告” 的 {task_name} 模块。 # 完成该模块必须需要以下数据： {task_need_data} # 有以下候选模型： {model_info} # 请判断： 1. 根据候选模型的输出，是否可以获得完成该模块需要的数据?（{task_need_data}） 2. 如果不可以，或者需要超过四个模型，那么请直接告诉我 ‘‘不可以’’。（例如，信息统计模型无法获取研发投入信息） 3. 如果可以，你应该如何选择和组合这些模型？请注意，最多只能选择四个模型，如果超过四个，参考第二点。（例如，信息统计模型可以获取专利数量信息） # 返回的格式有两种： - ‘‘不可以。因为无法获取 xx 信息” - ‘‘可以。流程是：1. 使用 xxx 进行 xxx\n2. 使用 xxx 进行 xxx”

表 6-6 工具调用与内容修正的大模型提示

Table 6-6 Prompts for the Tool Invocation and Content Correction Module

工具调用	### 任务 工具判断。判断以下是否存在符合条件的工具。如果没有，则直接回答‘无法完成’；如果有，则请从中选择合适的工具，用于完成当前需求。 ### 需求 {domain} 技术报告-{title} 中包含以下模块：{tool_list}，而其中你需要完成的是 {tool_key} 模块。 ### 工具 {tools_info} ### 要求 1. 选择的工具必须能够提供完成 {tool_key} 需求的全部或部分需要的数据，可以提供地不完整，但是不能完全无关； 2. 选择的时候需要考虑与其他模块间的重合度，根据情况可以舍弃重合度高的非关键工具； 3. 选择的时候务必带有客观性和严肃性，不要有情感倾向； 4. 请严格遵守工具输出的信息，请注意前后回答、逻辑的一致性； ### 返回格式 1. 需求分析。 2. 工具分析。包括分析该工具可提供什么信息？该信息和需求有关联吗？ 3. 综上所述，我的回答是：‘‘无需再选择” 或者综上所述，我的回答是：“xx 模型”
内容修正	# 任务 我将给你一段很长的信息文本，请理解背景并耐心阅读文本内容，然后以长文的形式进行信息的过滤与整合。 # 背景 {domain}{tool_key} # 信息： {text} # 要求 1. 保留的内容和 {domain}{tool_key} 相关；2. 以长文的形式进行总结，保留的内容尽可能丰富、细致；3. 内容尽可能流畅，不要有任何分点的格式；4. 不要过度忽略重要细节，重要细节可能对后续分析有极大帮助；5. 直接回答整合后的长文即可。

表 6-7 报告生成与优化的大模型提示

Table 6-7 Prompts for the Report Generation and Optimization Module

初始撰写	### 任务 我将给你一些对撰写 {domain}{tool_key} 有帮助的信息，它们可能比较长，请你耐心阅读并撰写 {domain} 行业技术分析报告--“{tool_key}”模块的内容。 ### 信息 {text} ### 要求 1. 第一步先简要梳理关键信息；2. 第二步重点结合 {tool_key}，对这些信息分析彼此联系；3. 最后，扩展思维，围绕 {tool_key} 进行技术层面的深入思考；4. 以“### {tool_key}’’ 开头，直接输出报告即可。
审核	以下是一段 {domain} 领域技术报告对 {subtitle} 相关的内容，请你围绕 {subtitle}，阅读并审查内容，返回作为专家 2-3 条最核心的批改意见。 ### 内容 “{agent_1}” 要求返回格式：“1. 意见 1；2. 意见 2；3. 意见 3”
优化	user: 初始撰写提示 assistant: {draft} user: {advice} 当然这只是我的意见，因为我并不了解原信息的全部内容，你可以根据自己的理解对初稿进行润色，使它质量更高、更流畅。 assistant:

附录 D

自动报告生成样例

在 2023 年 5 月 1 日至 2023 年 6 月 30 日期间，台湾积体电路制造股份有限公司、株式会社、福建省晋华集成电路有限公司在芯片制造领域发表的专利数量较多。以下是对这些龙头企业专利质量的分析：

台湾积体电路制造股份有限公司在半导体结构和半导体装置方面的专利质量较高。其专利涉及半导体结构的改进、半导体装置的设计以及半导体装置的形成方法。这些专利表明台湾积体电路制造股份有限公司在半导体领域的技术创新和研发实力较强，注重提升半导体器件的性能和可靠性。

株式会社在层叠基板、半导体封装和晶体振荡器方面的专利质量较高。其专利涉及层叠基板的设计、半导体封装的结构以及晶体振荡器的制造方法。这些专利显示了株式会社在芯片制造领域的技术创新和研发能力，特别是在基板设计和封装技术方面具有竞争优势。

福建省晋华集成电路有限公司在半导体器件及其制备方法、半导体器件及其形成方法以及半导体存储装置及其制作工艺方面的专利质量较高。其专利涉及半导体器件的设计、制备方法以及存储装置的制作工艺。这些专利表明福建省晋华集成电路有限公司在半导体领域的技术创新和研发实力，尤其在半导体器件和存储装置方面具有一定的竞争力。

综合来看，这些龙头企业专利质量较高，表明他们在芯片制造领域的技术创新和研发实力较强。他们的专利主要涉及半导体结构、装置、封装和存储等领域，这些技术是芯片制造行业的重要发展方向。

进一步思考，龙头企业专利质量的提升离不开他们对研发的持续投入和技术创新能力的培养。他们在关键技术领域的专利布局较为严密，通过不断创新和优化产品设计，提升产品的性能和可靠性。此外，龙头企业还应关注专利质量的持续提升，加强对专利的法律保护力度，以保护自身的技术成果和市场竞争力。

在未来，芯片制造行业将继续发展，龙头企业应继续关注技术创新和专利质量的提升，以保持竞争优势。同时，他们也应加强与其他企业的合作和技术交流，共同推动行业的发展。

半模板式

在2023年5月1日至2023年6月30日期间，芯片制造领域的专利活动活跃，其中台湾积体电路制造股份有限公司、株式会社以及福建省晋华集成电路有限公司的专利发表数量较多。以下是对这些公司专利质量的详细分析。

首先，从专利类型分布来看，台湾积体电路制造股份有限公司的发明公开类专利占比最高，达到78.6%，表明该公司在技术创新和公开分享方面具有较强的意愿和能力。而株式会社和福建省晋华集成电路有限公司的发明公开类专利占比也较高，分别为78.6%和71.4%。在实用新型类专利方面，福建省晋华集成电路有限公司的占比最高，达到14.3%，这可能反映了该公司在技术应用和产品改进方面的需求。

从IPC分布来看，这三个公司的专利主要集中在半导体器件及其制造方法、半导体结构、晶体振荡器、存储装置等领域。这显示出芯片制造领域的研究热点和技术发展趋势，也与当前市场需求和技术进步紧密相关。

在专利分数分布方面，台湾积体电路制造股份有限公司和株式会社的专利分数较高，分别有21.4%和15.4%的专利分数在60-70分，这表明这两家公司在芯片制造领域的技术创新和专利质量上具有较高的竞争力。而福建省晋华集成电路有限公司在专利质量上仍有提升空间，其60-70分的专利分数占比与株式会社相当。

具体到专利描述，台湾积体电路制造股份有限公司、株式会社和福建省晋华集成电路有限公司的专利质量表现各异。台湾积体电路制造股份有限公司的专利质量较高，涉及半导体结构、半导体装置及其形成方法等领域；株式会社的专利质量较高，涉及层叠基板、电子设备的制造方法、晶体振荡器及晶体振荡器的制造方法等领域；福建省晋华集成电路有限公司的专利质量也有不错表现，涉及半导体器件及其制备方法、半导体器件及其形成方法以及半导体存储装置及其制作工艺等领域。

综合来看，这三个公司在芯片制造领域的专利质量表现各异。台湾积体电路制造股份有限公司和株式会社在专利数量和质量上均表现突出，是行业内的佼佼者。福建省晋华集成电路有限公司在专利数量上也有不错的表现，但在专利质量上仍有提升空间。这些公司的专利主要集中在半导体结构、半导体装置、晶体振荡器、存储装置等方面，显示出芯片制造领域的研究热点和技术发展趋势。

在技术层面，我们可以进一步思考以下几个方面：
1. 台湾积体电路制造股份有限公司、株式会社和福建省晋华集成电路有限公司在专利质量上的差异可能与其技术创新能力、研发投入和战略定位等因素有关。如何通过提升专利质量来增强公司的技术竞争力和市场地位，是这些公司需要深入思考的问题。
2. 芯片制造领域的专利集中在半导体结构、半导体装置、晶体振荡器、存储装置等方面，这些领域的技术创新和市场竞争可能对公司的专利质量和研发方向产生重要影响。如何把握技术发展趋势和市场需求，优化专利布局和研发策略，是这些公司需要关注的关键问题。
3. 专利质量的提升不仅需要公司在技术创新和研发投入上加大力度，还需要在知识产权管理和保护方面做好工作。如何建立健全的知识产权管理体系，提高专利申请和维护的效率和质量，是这些公司需要重视的问题。

综上所述，龙头企业专利质量的高低与其技术创新能力、研发投入和战略定位等因素密切相关。在芯片制造领域，台湾积体电路制造股份有限公司和株式会社在专利数量和质量上表现突出，而福建省晋华集成电路有限公司在专利数量上也有不错的表现，但在专利质量上仍有提升空间。这些公司的专利主要集中在半导体结构、半导体装置、晶体振荡器、存储装置等方面，显示出芯片制造领域的研究热点和技术发展趋势。在技术层面，公司需要关注技术创新、市场需求、知识产权管理等方面的问题，以提升专利质量和增强技术竞争力。

AutoRG

图 6-1 芯片制造领域下的报告样例对比：龙头企业专利质量分析

Fig. 6-1 Report Comparison in Chip Manufacturing: Patent Quality Analysis of Leading Enterprises

3D打印技术近年来发展迅速，从原材料的创新，到核心硬件的改进，再到辅助运行系统的优化，以及设备制造和应用的拓展，都体现了这一技术领域的活力和创新力。

在3D打印原材料方面，专利聚焦于提高材料性能和打印质量，如实现光固化打印机自动调平的方法及存储介质，这表明行业正在努力提升原材料对打印过程的适应性和打印效率。同时，对鞋模金属粉末制备方法及3D打印方法的探索，显示出在精细材料处理和应用方面的进步。

核心硬件方面的专利集中在提高设备性能和稳定性，如提高投影设备曝光稳定性的图像处理方法，以及采用平行填充线和预设分割点的多激光器扫描方法。这些技术的发展，意味着3D打印硬件正朝着更高效、更精确的方向发展。

在辅助运行系统方面，连续自动运行方法及运行系统、3D印刷辅助机构的提供，显示了3D打印在自动化和智能化方面的进步。这些改进有助于提升生产效率，降低操作难度，使3D打印技术更加易于普及和应用。

在设备制造领域，主动保温曲线和加热装置的打印件成型方法、立体成型设备的设计，都体现了在提升打印质量和效率方面的探索。这些创新有助于降低生产成本，提高生产效率，进一步推动3D打印技术在制造业的应用。

在3D打印应用方面，根据树脂材料和平台高度控制树脂材料打印的方法、提高打印效率的三维模型打印方法等，都显示出行业在打印技术应用方面的深入研究。这不仅提高了打印质量，也扩大了3D打印技术的应用范围。

最后，在3D打印服务平台方面，双侧面补偿机构的3D打印设备、自动调平3D打印机托盘平台等，显示出平台在提供稳定、高效打印服务方面的创新。这些服务平台的发展，有助于推动3D打印技术的普及和应用。

综合以上分析，3D打印技术在各个子行业都呈现出积极的创新态势，特别是在原材料创新、硬件性能提升、自动化和智能化方面。未来，随着技术的进一步发展，3D打印技术在制造业、生物医学、航空航天等领域的应用将更加广泛，为各行各业带来更多的创新和变革。

半模板式

在2023年第一季度的3D打印行业专利分析中，我们观察到了几个显著的技术发展趋势。专利类型的分布显示，发明公开类专利的占比领先，这不仅表明了3D打印技术研发的创新性和开放性，也反映了行业对新技术披露的积极态度。实用新型类专利的较高比重揭示了技术实用化的追求，可能是对市场需求的快速响应。尽管发明授权类专利的比例相对较低，这可能反映了从研发到商业化的转化难度，但它也指向了技术成熟度和商业化潜力的一些挑战。外观设计类专利的存在，尽管占比最低，但仍显示出3D打印产品在外观创新上的努力。

专利评分分布的分析揭示了3D打印领域的专利质量较高，尤其是高分段专利的较高占比，这表明了研发投入的显著性和技术水平的先进性。在IPC分布上，增材制造设备及其相关零件和附件的专利布局占据了核心位置，凸显了行业对设备优化的关注。同时，增材制造的过程和数据处理方面的专利数量也较为突出，这反映了行业在提高制造效率和精度方面的不断探索。

进一步的聚类分析揭示了3D打印技术发展的多个方向。在原材料方面，专利技术的进步正在提升原材料的性能，增强打印控制的成功率。核心硬件的创新，如个性化打印方法和提高曝光稳定性的图像处理技术，正在推动设备性能的提升。辅助运行技术的进步，如多设备自动运行和电子测绘技术的结合，正在提高3D打印的高效运行。设备制造方面的专利揭示了在打印件成型方法和三维虚拟形象生成上的技术探索。应用方面的专利指向了根据特定材料和平台高度控制打印的方法，以及提高打印效率的三维模型打印技术。服务平台方面的专利集中在提高3D打印设备的稳定性和解决方案的效率上。

综合以上分析，3D打印子行业正在经历技术多元化与应用高效化的转变。技术创新的焦点集中在提升设备性能、优化打印过程、改进原材料特性和提高服务平台效率上。未来技术发展可能围绕以下几个方向：一是持续提高3D打印硬件性能，探索更高效、更稳定的设备；二是研发新型原材料，以提高打印质量和效率；三是深化数据处理和自动化技术，实现更智能的3D打印；四是拓展3D打印在更多领域的应用，如生物医学、建筑等。通过这些方向的努力，3D打印技术有望实现更广泛的应用和更深层次的影响。

AutoRG

图 6-2 3D 打印领域下的报告样例对比：子行业技术动态分析

Fig. 6-2 Report Comparison in 3D Printing: Analysis of Technological Dynamics in Sub-industries

学位论文数据集

表 1.1 数据集页

关键词 *	密级 *	中图分类号	UDC	论文资助
大语言模型；协同推理；小型专家模型；大模型推理增强；智能体	公开			
学位授予单位名称 *		学位授予单位代码 *	学位类别 *	学位级别 *
北京交通大学		10004	工学	硕士
论文题名 *		并列题名		论文语种 *
融合小型专家模型的大语言模型推理增强研究				中文
作者姓名 *	张京		学号 *	22120464
培养单位名称 *		培养单位代码 *	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直门外上园村 3 号	100044
学科专业 *		研究方向 *	学制 *	学位授予年 *
计算机科学与技术		自然语言处理	三年	2025
论文提交日期 *	2025 年 6 月			
导师姓名 *	桑基韬		职称 *	教授
评阅人	答辩委员会主席		答辩委员会成员	
	贾彩燕		黄华 王博	
电子版论文提交格式 文本（✓）图像 () 视频 () 音频 () 多媒体 () 其他 () 推荐格式：application/msword; application/pdf				
电子版论文出版（发布）者		电子版论文出版（发布）地		权限声明
论文总页数 *	78			
共 33 项，其中带 * 为必填数据，为 21 项。				