模型互联中多模型串并联协作推理

刘忠仁,李哲涛*,王建辉,肖 勇,曾曦玉,李 俊,莫光峰
(暨南大学信息科学技术学院,广东广州 511346)

摘 要: 大语言模型(Large Language Models,LLMs)凭借其庞大的参数规模和强大的语义表达能力,在自然语言处理、计算机视觉等领域取得突破性进展,并逐渐成为智能系统的关键基础. 然而,随着模型轻量化、本地化定制及场景专用化需求持续增强,面向特定任务开发的专有化模型快速涌现. 这类模型通常在局部领域具备能力优势,但难以独立覆盖多任务、多领域的复杂推理需求,从而推动了多模型协作推理的研究. 现有研究多侧重于模型融合或单一协作范式,难以充分挖掘各模型间的优势互补潜力,且在协作结构和路径机制方面缺乏系统性的探索. 为此,本文提出一种面向模型互联场景的多模型协作结构推理方法,构建了由线性链式结构向多路径组合结构演进的协作推理体系. 在基础协作层面,设计了串联推理(Serial Inference,SI)与并联推理(Parallel Inference,PI)两种核心范式,分别通过阶段式信息传递与多模型并行处理增强推理过程中的语义收敛性与信息覆盖度. 在此基础上,进一步从协作范式层面提出了"先串后并"(Serial-to-Parallel,S2P)与"先并后串"(Parallel-to-Serial,P2S)两种组合策略,实现协作路径在深度与广度之间的动态调度,拓展了多模型协作的结构表达能力与推理能力边界. 本文在数学推理、知识理解和符号推理三类典型任务上搭建了系统实验框架,对四类协作策略进行了全面评估. 实验结果表明,四类协作策略相较于单模型推理在平均准确率上分别提升了24.33、16.66、26.66和25.33个百分点. 进一步分析发现,组合协作策略在融合串联与并联结构优势的同时,能够有效压缩整体推理时延,并在相较于最优单模型可接受的时延增量条件下,实现了更高的推理准确率,展现出在多任务场景下更优的性能-效率的权衡. 此外,本文还系统分析了不同模型路径配置在协作过程中的表现差异,为多模型组网结构设计、协作机制优化及大规模模型互联体系的构建提供了理论依据与实证支撑.

关键词: 大模型;模型互联;多模型协作;串联协作;并联协作;组合推理
基金项目: 国家自然科学基金(No.W2411053,No.U23B2027)
中图分类号: TN92 **文献标识码:** A **文章编号:** 0372-2112(2025)11-3817-19
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250503

Multi-Model Serial and Parallel Collaborative Inference in AI-ModelNet

LIU Zhong-ren, LI Zhe-tao*, WANG Jian-hui, XIAO Yong, ZENG Xi-yu, LI Jun, MO Guang-feng
(*College of Information Science and Technology*, *Jinan University*, *Guangzhou*, *Guangdong* 511346, *China*)

Abstract: Large language models (LLMs), empowered by massive parameter scales and strong semantic representation capabilities, have achieved breakthrough progress in natural language processing, computer vision, and related fields, and have gradually become a key foundation of modern intelligent systems. However, increasing demands for lightweight deployment, on-device customization, and scenario-specific specialization have led to the rapid emergence of task-specific models. Although these specialized models exhibit strong capabilities within their respective domains, they are insufficient for handling complex multi-task and multi-domain reasoning independently, which motivates research on multi-model collaborative inference. Existing studies primarily focus on model fusion or single collaboration paradigms, which limits the exploitation of complementary strengths across models and lacks systematic exploration of collaboration structures and path mechanisms. To address these challenges, this study proposes a collaborative inference framework for model-interconnection scenarios, enabling an evolutionary shift from linear chain structures to multi-path composite structures. The framework formalizes two basic paradigms—serial inference (SI) and parallel inference (PI)—and further introduces two hybrid strategies, serial-to-parallel (S2P) and parallel-to-serial (P2S), to dynamically coordinate depth- and breadth-oriented collaboration pathways. Comprehensive experiments on mathematical reasoning, knowledge understanding, and symbolic reason-

收稿日期:2025-06-11;录用日期:2025-11-11;责任编辑:覃怀银
*通讯作者:李哲涛

ing show that SI, PI, S2P, and P2S improve accuracy by 24.33, 16.66, 26.66, and 25.33 percentage points, respectively, compared with single-model inference. Additional analysis shows that hybrid collaboration significantly reduces overall inference latency while achieving higher accuracy, demonstrating a superior performance-efficiency trade-off. Moreover, the study reveals the structural impacts of different collaboration paths, offering theoretical insights and empirical evidence for the design of multi-model networks and efficient model-interconnection systems.

Key words: large model; model interconnection; multi-model collaboration; serial inference; parallel inference; composite collaborative inference

Foundation Item(s): National Natural Science Foundation of China (No.W2411053, No.U23B2027)

1 引言

随着深度学习技术的快速发展,大语言模型(Large Language Models,LLMs)已广泛应用于自然语言处理、计算机视觉以及多模态任务等领域,成为推动人工智能进步的关键动力[1~3]. 近年来,随着对本地部署与任务定制化的需求不断增强,通用大模型在特定场景中的适应性受限,难以满足在应用场景中的性能要求[4]. 这一趋势促使了各领域开始构建面向教育[5]、网络安全[6]、医疗[7]等垂直领域的专有化模型,形成了模型能力呈现明显领域差异、部署形式多样的格局. 然而,由于模型间知识结构、表达能力及推理机制存在差异,单一模型在处理复杂或跨域任务时往往面临信息覆盖不足与功能局限等问题[8,9]. 在此背景下,异构化模型因缺乏统一协作机制而形成“知识孤岛”问题(图1),严重制约了智能系统整体效能的发挥. 因此,亟需构建支持模型间协作与能力互补的高效推理机制,以打通模型间的信息通道,推动模型互联互通成为智能系统演进的重要趋势.

图1 私有化模型时代的知识孤岛与协作困境

当前,多模型异构部署逐步成为智能系统架构的重要特征,模型间的互联能力已成为支撑复杂任务协同处理的重要基础[10,11]. 通过统一的接口协议与调度机制,模型互联实现了模型间的信息传递与功能协同,为进一步的协作推理提供了底层支撑. 在此基础上,模型协作推理[12]作为实现知识共享与能力融合的关键途径,正发展为人工智能系统构建的重要研究方向.

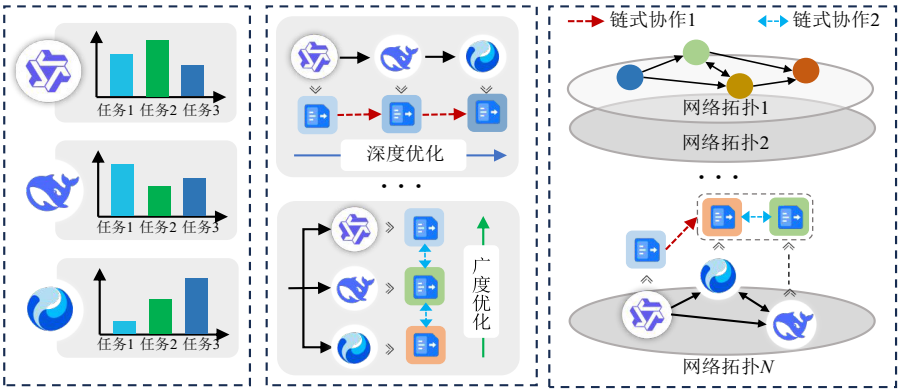
现有的模型协作推理方法主要包括两类:一类是基于知识迁移或参数集成的模型融合策略,通过整合多个模型的参数与语义能力,构建统一模型以提升整体性能,但此类方法通常训练成本高,且部署灵活性差[13,14];另一类是基于模型集成的决策融合方法,通过预定义策略对多个模型的预测结果进行选择与融合[15,16]. 尽管此类方法在一定程度上提升了模型的推理表现,但其协作路径通常静态且单一,缺乏对协作结构与推理流程的深入优化,限制了模型间优势互补潜力的充分释放.

针对上述问题,本文面向模型互联应用场景,提出了一种支持结构演化与路径组合的多模型协作推理框架,如图2所示. 该框架在方法设计上引入由线性链式结构向多路径组合结构演进的协作思路,系统构建了串联协作推理(Serial Inference,SI)与并联协作推理(Parallel Inference,PI)两类基础协作结构. 其中,SI通过层级式信息传递实现模型间的逐步决策优化;PI则通过多模型同步推理与结果整合,提升信息覆盖广度与系统的容错冗余能力. 在此基础上,本文进一步提出了先串后并(Serial-to-Parallel,S2P)与先并后串(Parallel-to-Serial,P2S)两种结构组合策略,构建了可适配不同任务特性的协作路径空间. 该组合策略不仅引入了推理链路的可组合结构与协作顺序的动态调度机制,还在信息路径与语义融合层面实现了更高维度的结构表达. 整体方法体现了从“点式推理”向“链式优化”,再到“结构化协同”的系统演化,为构建具备灵活性、泛化性与鲁棒性的模型协作体系提供了新的方法范式.

基于上述背景,本文旨在探索大模型互联互通条件下的协作机制,研究多模型间的协作方式及其对推理性能的影响,揭示多模型协作推理的结构特性与性能优势,为构建高效、可扩展的模型互联架构提供理论依据与实践参考. 本文的主要贡献如下.

(1)围绕模型互联场景下的协同推理需求,本文提出并系统化定义了SI与PI两类基础协作范式,从结构建模与机制设计两个维度分析其协作过程中的语义传递特性、路径组织模式及任务适用性,为多模型协作体系奠定了结构清晰、机制明确的基础框架.

(2)在基础协作范式之上,进一步提出S2P与P2S



(a) 单点推理　　(b) 链式协作推理　　(c) 网络式协作推理

图2　支持结构演化与路径组合的多模型协作推理框架示意图

两种组合协作策略,实现协作结构从局部单一范式向全局多路径组合空间的扩展. 该设计在兼顾协作深度与信息广度的基础上,有效提升了协作体系在任务多样性与结构复杂性条件下的适应性与稳定性.

(3)构建了一个包含4类协作结构与24种模型路径的实验评估框架,并采用准确率和时延2项指标,以系统评估所提协作策略在GSM8K、CEVAL和Mathematics任务上的性能表现. 实验结果显示,在多任务场景下,组合协作推理不仅显著提升了准确率,还相较于SI与PI结构展现出更优的延迟控制能力. 尽管延迟略高于最优单模型,但以可接受的延迟增量换取了显著的准确率提升,展现出更稳定且更具性价比的性能-效率协同优势.

2　相关工作

2.1　大模型推理

近年来,生成式LLMs在人工智能领域取得了突破性进展,其应用范围已从传统的序列建模拓展至复杂的语言理解与生成任务,成为推动人工智能技术演进的重要驱动力. Transformer架构[17]的提出及其自注意力机制的引入,有效缓解了长距离依赖建模的难题,为构建高性能语言模型提供了理论基础与结构支撑. 在此背景下,涌现了一系列性能强大的LLMs,开启了"百模竞技"的局面. 以GPT系列[18]和LLaMA系列[19]为代表的闭源模型在多个基准任务中取得领先性能,推动了大模型技术的快速成熟. 与此同时,诸如DeepSeek[20]、Qwen[21]、Kimi[22]等开源模型在参数规模、训练策略和架构设计等方面持续发展,展现出优异的性能与良好的可扩展性. 随着大模型在自动驾驶[23]、智能助手[24]、内容生成[25]与医疗辅助诊断[26]等领域的深入应用,人工智能逐步从单点任务向多模态、跨领域综合智能形态演进.

在大模型的推理过程中,其核心机制主要依赖于Transformer架构中自注意力机制和前馈网络的协同作用. 具体而言,设输入序列为$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,每个输入词$\boldsymbol{x}_n$经过嵌入映射后,利用线性变换分别生成查询($\boldsymbol{Q}$)、键($\boldsymbol{K}$)、值($\boldsymbol{V}$)向量,并计算其自注意力分数:

$$\text{Attention}\{\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\}=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

其中,d_k为键的维度,用于对点积注意力缩放以确保梯度稳定. 随后,注意力输出与原始输入通过残差连接并经层归一化处理,结果再输入至前馈网络(Feed-Forward Network,FFN). 该网络由两个线性变换和非线性激活函数构成,其形式表示为

$$\text{FFN}(x)=\max(0,\boldsymbol{x}\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{2}$$

其中,$\boldsymbol{W}_1$、$\boldsymbol{W}_2$为线性变换矩阵;$\boldsymbol{b}_1$、$\boldsymbol{b}_2$为相应的偏置项,非线性函数$\max(0,\cdot)$则是为了引入非线性模型,增强模型的表达能力与特征变换能力. 在推理阶段,大模型通常采用自回归生成机制,即通过逐步预测下一个词的概率分布来生成完整的文本,其概率模型可表达为

$$P(\boldsymbol{x}_{t+1}|\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t)=\text{softmax}(\boldsymbol{W}_{\text{out}}\boldsymbol{h}_t+\boldsymbol{b}_{\text{out}}) \tag{3}$$

通过自回归生成机制,大模型逐步处理输入词元,并在每一步生成下一个词元的概率分布. 在每个迭代中,新生成的词元会追加至当前输入序列前缀,形成新的上下文表示,并作为模型输入参与下一步预测,直至触发结束符或达到最大输出长度. 该生成过程依赖于模型在训练阶段所使用的大规模语料库,模型通过对语言结构与语义模式的深入学习,具备了较强的内容理解与生成能力.

值得注意的是,不同大模型由于训练数据构成、预训练目标函数以及模型架构设计的差异,在知识覆盖范围、表述特征和任务适应能力等方面表现出显著的异质性. 例如,部分模型在法律、医疗等专业领域具备较强的术语解析与推理能力,而另一些模型则更擅长创意写作、对话生成等开放性任务. 这种模型能力的差

异性使得大模型在实际应用中具备互补性，为构建多模型协同推理体系提供了重要基础.

2.2 模型协作推理

模型协作推理旨在整合多个模型在表示能力、知识结构与任务特长方面的互补优势，以提升复杂推理任务中的泛化能力与鲁棒性. 该方法能够有效缓解单一模型在多任务场景中可能出现的过拟合、表示偏差等问题，近年来已在推荐系统[27,28]、医学问诊[29,30]、云边协同[31,32]等应用领域得到广泛探索. 例如，Li等人[33]提出通过可信与不可信模型的协作机制，有效缓解了大模型在版权审查、数据污染与隐私风险方面的脆弱性；Hoang等人[34]则将传统机器学习翻译模型与大语言模型集成，在多个基准任务上实现了显著性能提升. 现有多模型协作推理的研究路径主要可分为两大类：一类侧重于结构整合的模型融合方法，另一类则聚焦于推理过程协同的模型集成策略.

在模型融合方面，Wan等人[35]通过多教师知识蒸馏机制，将多个预训练大模型的结构知识迁移至单一目标模型，在多项任务中取得了性能显著提升. Bansal等人[36]提出了跨模型注意力机制，在融合模型特征表达能力的同时增强了模型对未知任务的泛化能力. 尽管此类方法可构建独立且高性能的单一模型系统，但普遍存在训练代价高、结构固化与更新不灵活等问题，难以适应动态、多变的在线推理需求.

相较而言，模型集成方法更强调模型间的信息流动与知识共享，通常通过并行生成、阶段交互或职责分配等机制，实现多模型协同推理与能力互补. 例如，Ven等人[37]通过在不同阶段引入多个模型共同参与故事创作，构建了首个多模型协作生成的故事数据集CollabStory，展示了协作推理在创造性语言任务中的潜力. Meta AI提出的Collaborative Reasoner（Coral）框架[38]通过基于自对弈的合成对话机制，使多个语言模型实现持续协同与自我改进，在开放式生成任务中表现突出；MoSA框架[39]利用多模型协作搜索范式，显著提升了大模型在复杂问题求解过程中的解空间覆盖能力与鲁棒性. 同样地，在更细粒度的协作机制上，Wang等人[40]在模型互联背景下提出了一种基于Token级协作的推理架构DuetNet，旨在通过多模型推理共识的聚合机制，有效降低错误推理路径选择的风险；Yu等人[41]实现了基于词元级预测分布的对齐与集成策略，有效降低了早期错误传播对整体推理结果的影响；Xu等人[42]设计了基于困惑度评估的跨度选择机制，使多个候选模型在每一生成阶段并行输出多个内容单元，并通过打分机制筛选最优结果，有效提高了生成稳定性与表达质量. 此外，级联推理[43,44]与投机采样机制[45,46]通过主从式模型协作结构，将低容量模型作为快速预估器，高容量模型作为校验者，实现了在推理效率与输出准确性之间的有效权衡.

尽管现有研究在协作策略设计、粒度控制与应用扩展等方面取得了重要进展，但大多数方法仍停留在模型间协作范式的构建，缺乏对更高层次协同结构的系统拓展. 特别是在协作结构从线性耦合向多路径组合的演进过程中，尚缺乏对其路径设计、执行机制与性能潜力的深入挖掘，制约了协作体系在结构灵活性、协同鲁棒性与任务泛化能力等方面的进一步提升.

3 串并联协作推理

基于对LLMs推理机制的深入分析，本文首先构建了SI与PI两种基础协作范式，并在此基础上提出了S2P和P2S两种组合协作策略，用于融合串联与并联结构优势，实现更高效、更灵活的协同推理能力.

3.1 串联协作推理

在多模型协作推理框架中，SI是一种结构明确、信息流动具有严格顺序依赖性的推理范式. 其基本思想如下：将多个异构模型按照前后依赖关系依次连接形成链式结构，后续模型以前一阶段模型的输出或提示信息作为输入，逐步对语义信息进行抽象、约束与增强，最终完成任务预测. 因此，串联协作推理可形式化一个有向信息传递链路：

$$\mathrm{Node}_1\left(\boldsymbol{x}^{(1)},\boldsymbol{y}^{(1)}\right)\rightarrow\mathrm{Node}_2\left(\boldsymbol{x}^{(2)},\boldsymbol{y}^{(2)}\right)\rightarrow\cdots\rightarrow\mathrm{Node}_L\left(\boldsymbol{x}^{(L)},\boldsymbol{y}^{(L)}\right) \tag{4}$$

其中，$\mathrm{Node}_L(\cdot)$表示第L个模型节点，接收输入$\boldsymbol{x}^{(L)}$，输出预测$\boldsymbol{y}^{(L)}$，并将中间语义结果或阶段性解答以提示的形式传递给后续节点. 该协作范式可定义如下.

定义1 串联协作推理. 设存在一个有序模型序列$\boldsymbol{M}=\{M_1,M_2,\cdots,M_K\}$，其中每个模型$M_k$为映射函数$M_k$：$\mathcal{Z}_{k-1}\rightarrow\mathcal{Z}_k$，其中$\mathcal{Z}_0$表示原始输入空间，$\mathcal{Z}_K$表示最终输出空间. 对于模型输入$\boldsymbol{x}\in\mathcal{X}$，串联推理过程可描述为

$$\boldsymbol{y}=\boldsymbol{z}_K=M_K\left(M_{K-1}\left(\cdots M_1(\boldsymbol{x};\boldsymbol{p}_1)\cdots\right);\boldsymbol{p}_K\right) \tag{5}$$

其中，$M_K(\cdot)$表示模型推理生成；$\boldsymbol{z}_i$为模型M_i的输出；$\boldsymbol{p}_K$表示模型K的专属提示词（如角色设定或语境约束），最终输出结果为$\boldsymbol{y}\in\mathcal{Y}$.

这一结构设计受到人类认知中多步推理与链式思维机制的启发. 如图3所示，在处理诸如常识推理、结构化问答等任务时，人类往往倾向于分阶段思考，通过层层递进的逻辑推演逐步逼近问题解. 串联协作模型正是对这一认知过程的模拟，通过在模型间建立顺序连接，使每一阶段能够引入更高层次、更具约束力的语义信息，从而提升整体推理的精度与一致性.

基于上述结构，串联协作推理的执行流程可描述如下：（1）当任务x发起服务请求后，首个模型节

点负责进行问题解析与初步语义建模,生成中间表示;(2)后续模型依次对前一阶段结果进行细化处理,逐层完成逻辑构建、信息补全与语义集成;(3)最后一级模型完成决策整合并输出最终答案. 该结构能够增强推理语义在模型序列中的连续性与一致性,同时为复杂任务提供了较为清晰的分阶段求解路径,并在一定程度上实现了从问题建模到答案生成的过程解耦与优化.

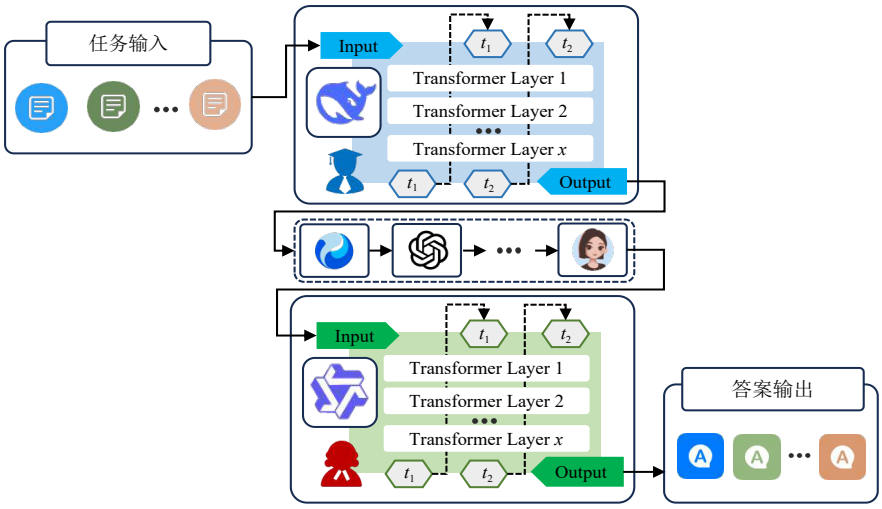


图3　串联协作推理示意图

值得注意的是,在串联协作推理结构中,各模型通过显式的链式顺序构成完整的推理链路,且每一阶段模型的输入均高度依赖于前一阶段模型的输出. 这一结构赋予推理流程明确的阶段性与递进性,但同时也引入了显著的链路依赖性:前序模型生成的中间表示不仅决定了后续模型的输入语境,也可能影响整个推理路径的语义方向与信息质量. 因此,本文针对串联协作推理提出如下顺序敏感性假设.

假设 1　顺序敏感性. 在串联协作路径中,前序模型输出的语义表达质量越低,对后序模型推理造成的扰动效益越大;该扰动将随模型链路的加深逐步放大,最终显著削弱整体推理性能的稳定性与一致性.

该假设表明,在缺乏显式纠错机制或反馈策略的条件下,串联结构可能存在一定的结构性脆弱性. 由于串联协作范式中的顺序依赖不仅决定信息传递的方向,也直接构成影响推理稳定性的关键路径,链路早期误差极易在后续阶段被逐级放大. 因此,串联协作推理在设计与应用时需特别关注模型排列顺序及路径选择,以减轻链路误差的累积效应.

3.2　并联协作推理

在多模型协作推理中,除了串联结构通过语义层级深化实现链式优化外,并联结构则提供了一种更具容错能力与结构稳定性的协作范式. PI 旨在构建同层级多模型的横向推理路径,使各模型基于自身内部的知识结构并行解析输入,并在统一的融合规则下聚合得到最终预测结果.

不同于串联结构的纵向信息链路,并联结构在逻辑上构建的是横向多视角链路,其信息流形式可表述为

$$\boldsymbol{x} \Rightarrow \left\{M_1(\boldsymbol{x},t), M_2(\boldsymbol{x},t), \cdots, M_K(\boldsymbol{x},t)\right\} \Rightarrow \mathcal{F}(\cdot) \Rightarrow \boldsymbol{y} \tag{6}$$

其中,$M_K(\cdot)$表示第 K 个模型在时间步 t 上对输入 $\boldsymbol{x}$ 的预测分布;$\mathcal{F}(\cdot)$为多模型融合器;$\boldsymbol{y}$ 表示最终预测. 因此,该协作推理范式可定义如下.

定义 2　并联协作推理. 设存在一个异构模型集合 $\boldsymbol{M}=\{M_1, M_2, \cdots, M_K\}$,其中各模型在结构设计、参数初始化或训练语料上存在差异. 对于给定输入 $\boldsymbol{x} \in \mathcal{X}$,每个模型在解码时间步 t 独立计算其对候选词元 $v \in \mathcal{V}$ 的预测分布,并归一化得到概率分布 $\mathcal{P}_k^{(t)}[v]$. 随后,通过加权概率融合策略与贪婪解码策略,选取得到该时间步最优预测结果:

$$\boldsymbol{y}^{(t)} = \underset{v \in \mathcal{V}}{\operatorname{argmax}} \left(\sum_{k=1}^{K} \frac{1}{K} \cdot \mathcal{P}_K^{(t)}[v] \right) \tag{7}$$

其中,融合策略采用等权融合,以最大限度保留不同模型的语义多样性. 重复该过程,直到满足生成终止条件(如生成<EOS>或达到最大步数 T),最终得到完整输出序列如下所示:

$$\boldsymbol{y} = \left\{\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \cdots, \boldsymbol{y}^{(T)}\right\} \tag{8}$$

如图4所示,该机制本质上在语义空间中构建了一组并行的多视角路径,采用“前期多样性探索+后期共识决策”的协作策略. 在前期阶段,不同模型对输入任务进行独立的语义建模与解码,从多个角度充分保留

语义多样性与互补性;后期阶段,在融合策略上采用贪婪解码作为默认的集成策略,以最大化当前时间步所有模型共同支持的候选项的联合置信度为原则,从而增强多模型间的语义共识与决策一致性.

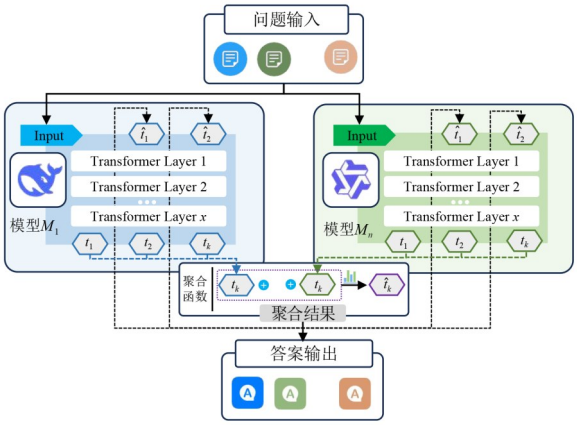


图4 并联协作推理示意图

并联推理充分利用了模型间的认知差异性,使不同模型能够从各自擅长的语义维度出发形成互补性表达,实现语义理解的多样性增强与结果生成的稳定性提升. 其推理流程包括:(1)各模型以相同的任务输入$\boldsymbol{x}$为条件,独立执行解码过程,生成阶段性词元预测分布;(2)每个解码步,融合模块对所有模型的预测结果进行联合评估,确定最优候选词元,并将其作为下一步解码上下文;(3)重复上述过程,至达到最大序列长度或生成终止符<EOS>. 因此,该范式在保持语义连贯性的同时,能够在一定程度上规避串行依赖导致的误差传播问题,展现出较强的稳定性与容错能力.

值得注意的是,为了保证预测路径的一致性和多模型协同输出的同步性,并联协作推理采用同步式融合机制,即在每一解码时间步,需等待所有模型完成该步预测,方可执行融合操作并生成下一个词元. 然而,这种机制在协作效率上引入了潜在的结构性延迟瓶颈:一旦某个模型在当前时间步的解码耗时显著增加,整个协作流程将被迫等待,使得并联结构的整体时延不再取决于模型的平均响应能力,而是受到动态最慢分支的单步响应主导. 基于此,本文针对并联结构的聚合特性提出如下假设.

假设2 最慢分支主导效应. 在同步聚合机制下,并联协作推理每一解码步骤的系统延迟由当前时间步响应最慢的模型分支动态决定. 由于各模型的推理延迟具有时间动态性,最慢分支在不同时间步不断变化,从而使整个并联协作路径呈现由“动态最慢分支”主导的延迟特性,并最终构成并联协作结构的效率瓶颈.

该假设指出,在同步聚合机制下,并联结构可能存在较为突出的结构性瓶颈:系统必须等待所有模型完成预测后方可执行词元级聚合,导致整体生成速度受限于最慢模型分支的响应能力,削弱了并行结构在理论上的吞吐优势. 因此,“最慢分支主导效应”成为影响并联协作推理延迟性能的核心瓶颈之一.

3.3 串并联组合协作推理

从协作推理的信息流动方式来看,SI和PI分别体现了两类互补的推理能力. 前者具备良好的阶段性建模特征,能够通过多级模型顺序协同实现语义的逐步精炼与推理空间的逐层收敛;而后者强调模型间的横向互补,通过多模型异构视角并行解读输入任务,增强信息覆盖的全面性和生成表达的鲁棒性.

基于上述观察,本文进一步提出了组合协作推理策略(Composite Collaborative Inference,CCI),在统一框架下集成SI与PI结构的互补优势. 该策略旨在“语义递进性”与“视角多样性”之间建立协同机制,使得信息在协作路径中既可获得逐层深化的建模优势,又能在关键阶段引入多模型的异构支持,从而提升整体推理过程的语义表达能力、稳定性与容错性能. 在该框架下,CCI可具体化为S2P组合推理和P2S组合推理这两种推理策略,如图5所示.

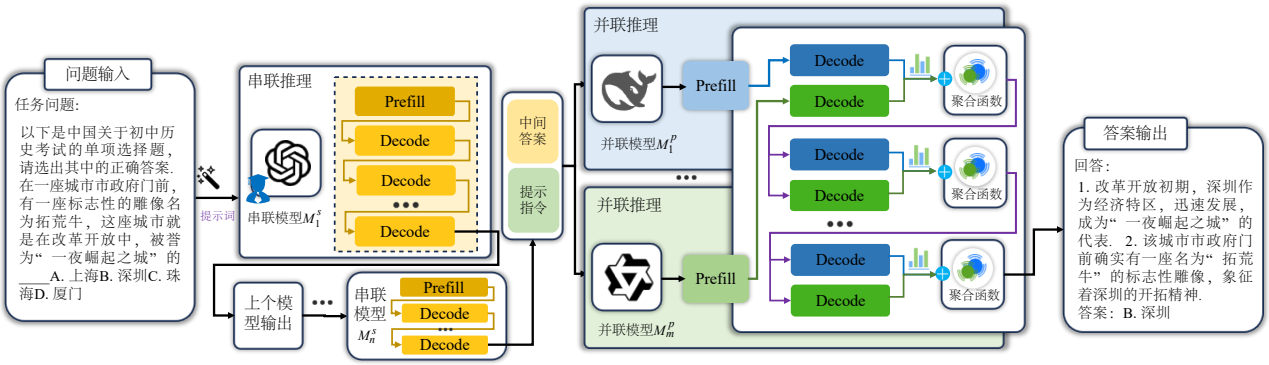
3.3.1 先串后并组合协作策略

S2P协作推理策略体现了一种“先构建主干、后引入多源补偿”的结构化协作思想,如图5(a)所示. 该策略在协作范式层面融合了串联结构的层级建模优势与并联结构的多视角补全能力:前阶段通过串联模型形成逐层递进的语义主干,使关键语义结构得到充分抽取;后阶段引入并联模型,以多角度的异构解读补充语义细节,增强生成过程的全面性与稳健性.

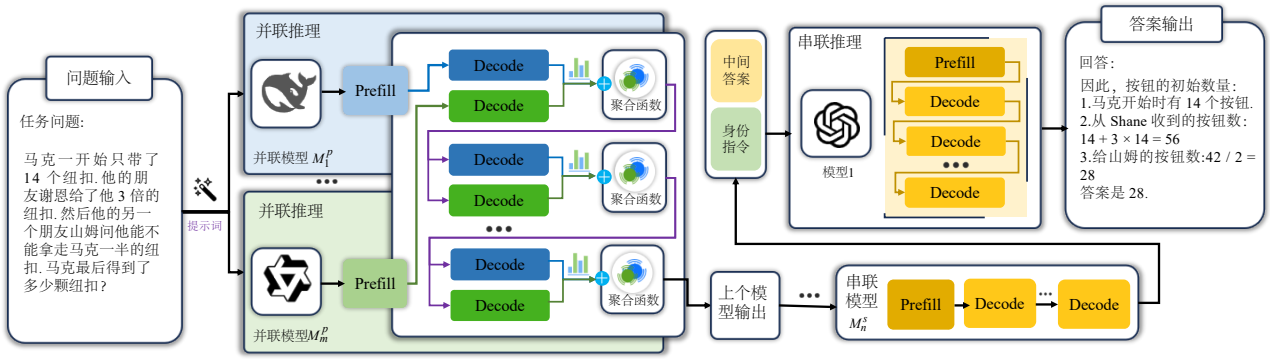
具体而言,S2P由串联子结构与并联子结构两阶段组成. 推理过程首先依赖串联模型序列逐步抽取输入中的核心语义线索,各级模型在接收前一阶段输出的基础上进行上下文重构与语义压缩,最终生成具备高度抽象性和上下文一致性的中间语义表示. 随后,该表示被同时输入至异构多模型构成的并联结构,各模型在共享上下文条件下独立解码当前词元,并通过融合机制聚合共识预测,形成整体序列生成. 因此,S2P的协作过程可定义如下.

定义3 S2P组合推理. 设串联模型序列为$\boldsymbol{S}=\{M_1^s,M_2^s,\cdots,M_K^s\}$,并联模型集合为$\boldsymbol{Q}=\{M_1^p,M_2^p,\cdots,M_J^p\}$. 当输入任务$\boldsymbol{x}\in\mathcal{X}$,串联结构逐层推理得到中间语义表示为$\boldsymbol{h}^{(K)}=M_K^s(\boldsymbol{h}^{(K-1)},\boldsymbol{p}^K)$. 该表示作为并联结构的共享上下文输入,各并联模型在第t个时间步预测词元分布为$p_t^{(j)}=M_j^p(\boldsymbol{z}_K)$,通过融合函数进行加权平均与贪婪聚合后,得到融合预测分布为

$$\boldsymbol{y}_t=\mathcal{F}(p_t^{(1)},p_t^{(2)},\cdots,p_t^{(J)}),\ t=1,2,\cdots,T \tag{9}$$



(*a*) S2P组合推理



(*b*) P2S组合推理

图5　串并联组合协作推理流程示意图

其中,$\mathcal{F}(\cdot)$表示贪婪解码的词元聚合函数. 根据大模型自回归推理机制,将融合后的词元作为当前步骤输出并用于更新生成序列,直到满足终止条件,最终形成输出序列$\boldsymbol{y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_T\}$.

S2P的推理流程如算法1所示,其核心机制是一种语义主干引导与多源融合补偿相结合的两阶段协作范式. 在串联阶段,模型序列构建自上而下的信息精炼路径,通过逐层语义抽象与结构压缩,提取具有全局一致性的中间语义表示;在并联阶段,该表示被同时输入至多个异构模型,各模型独立完成候选词元预测,融合模块则依据一致性最大化原则执行词元级集成,从而实现多角度语义的补偿与共识输出,提升语义建模与生成过程的整体质量.

基于该机制,本文在方法设计阶段提出如下任务适配性假设:对于结构化知识占主导,逻辑依赖显著且输入语义结构清晰的任务,S2P结构更适合先构建稳定的语义骨架,再通过多模型对关键要素进行冗余校验与多角度补偿,从而获得更高的生成准确性与稳健性. 例如,在CEVAL等常识问答类任务中,由于任务通常具备明确的输入结构与知识支撑,S2P可通过串联阶段完成问题建模与逻辑抽象,并在并联阶段引入语义扩展

算法1　S2P组合策略

输入:输入任务$\boldsymbol{x}\in\mathcal{X}$,串联模型序列$\boldsymbol{S}=\{M_1^s,M_2^s,\cdots,M_K^s\}$,并联模型集合$\boldsymbol{Q}=\{M_1^p,M_2^p,\cdots,M_J^p\}$,最大生成长度$T$,串联提示词$\boldsymbol{P}=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_K\}$

输出:生成的输出序列$\boldsymbol{y}$

1. 初始化输入表示:$\boldsymbol{z}_0\leftarrow\boldsymbol{x}$
2. FOR $k=1,2,\cdots,K$ DO
3. 　串联模型序列进行推理(如式5):$\boldsymbol{z}_k\leftarrow M_k^s(\boldsymbol{z}_{k-1},\boldsymbol{p}_k)$
4. END FOR
5. 初始化并联输入:$\boldsymbol{h}\leftarrow\boldsymbol{z}_K$
6. 初始化并联输出序列:$\boldsymbol{y}$
7. FOR　$t=1,2,\cdots,T$ DO
8. 　FOR　$j=1,2,\cdots,J$ DO
9. 　　并联集合预测第t个词元分布:$p_t^{(j)}\leftarrow M_j^p(\boldsymbol{h},\boldsymbol{y})$
10. 　END FOR
11. 　概率融合与贪婪解码选取词元(如式7):$\boldsymbol{y}_t\leftarrow\mathrm{Fuse}(p_t^{(J)},\cdots,p_t^{(1)})$
12. 　生成当前预测输出:$\boldsymbol{y}\leftarrow\boldsymbol{y}\oplus\boldsymbol{y}_t$
13. 　IF $\boldsymbol{y}$ == <EOS>　THEN
14. 　　BREAK
15. 　END IF
16. END FOR
17. RETURN $\boldsymbol{y}$

与结果修正机制，从而更好地契合此类任务的推理需求.

3.3.2 先并后串组合协作策略

P2S协作推理策略旨在构建一种“前期广泛感知、后期逻辑收敛”的信息处理路径，如图5(*b*)所示. 与强调由结构主干引导生成的S2P策略不同，P2S更侧重于在初始阶段保留模型之间的语义差异性和解读多样性，并通过后续的串联结构实现语义统一与逻辑闭环.

在推理流程上，P2S协作策略首先在前期阶段激活多个异构并联模型对共享输入任务并行建模，各模型基于自身的语义偏好生成候选词元的预测分布，并通过置信度融合机制在词元级形成联合表示，从而逐步构建中间语义序列. 随后，该表示作为共享上下文输入串联模型序列，逐层执行语义压缩与逻辑增强，最终生成结构化且收敛性强的预测结果. 基于该流程，P2S协作过程可定义如下.

定义4 P2S组合推理. 设输入任务$\boldsymbol{x}\in\mathcal{X}$，并联模型集合为$\boldsymbol{Q}=\{M_1^p,M_2^p,\cdots,M_J^p\}$，串联模型序列为$\boldsymbol{S}=\{M_1^s,M_2^s,\cdots,M_K^s\}$. 在每个生成时间步$t$，并联模型基于共享输入$\boldsymbol{x}$独立进行解码，分别生成该时间步候选词元的概率分布$\{p_t^{(1)},p_t^{(2)},\cdots,p_t^{(J)}\}$. 融合器对上述分布进行聚合，得到词元级联合表示分布$\boldsymbol{h}_t=\mathcal{F}(p_t^{(1)},p_t^{(2)},\cdots,p_t^{(J)})$，并将所有步生成的联合表示序列$\boldsymbol{h}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_T\}$作为中间表示，依次输入串联模型序列执行阶段性推理：

$$\boldsymbol{y}=M_K\left(M_{K-1}\left(\cdots M_1(\boldsymbol{h};\boldsymbol{p}_1)\cdots\right);\boldsymbol{p}_K\right)\tag{10}$$

其中，$\boldsymbol{p}_K$表示模型M_K^s的专属提示词，最终输出结果为$\boldsymbol{y}$.

P2S推理流程如算法2所示，其本质是一种“语义发散探索-结构收敛整合”的协同推理路径. 前期的并联阶段引导多角度建模与语义多样性生成，以缓解单路径推理中可能出现的早期偏倚风险；后期的串联阶段则逐级压缩冗余信息，聚焦关键逻辑路径，从而增强最终输出结果的一致性与整体鲁棒性.

基于P2S协作策略的机制特性，本文提出如下任务适配性假设：对于以逻辑推理主导、推理路径具有多解性且解空间开放度较高的任务类型，P2S结构更具优势. 其核心在于，前期的多路径语义发散能够覆盖更大范围的潜在解空间，而后续串联阶段的逐级收敛过程则有助于抑制语义噪声、强化主干逻辑，从而有效缓解早期误导带来的误差级联放大效应. 以GSM8K等数学逻辑类任务为例，P2S策略可通过并联模型并行探索多种解题路径，再由串联结构逐层筛选并聚焦关键步骤，构建符合因果逻辑的解题链条，从而提升推理闭环一致性与结果的逻辑正确性.

算法2 P2S组合策略

输入：输入任务$\boldsymbol{x}\in\mathcal{X}$，并联模型集合$\boldsymbol{Q}=\{M_1^p,M_2^p,\cdots,M_J^p\}$，串联模型序列$\boldsymbol{S}=\{M_1^s,M_2^s,\cdots,M_K^s\}$，最大生成长度$T$，串联提示词$\boldsymbol{P}=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_K\}$

输出：生成的输出序列$\boldsymbol{y}$

初始化并联输入：$\boldsymbol{x}_0\leftarrow\boldsymbol{x}$;
初始化中间表示序列：$\boldsymbol{h}$
FOR $t=1,2,\cdots,T$ DO
  FOR $j=1,2,\cdots,J$ DO
    并联集合预测第t个词元分布：$p_t^{(j)}\leftarrow M_j^p(\boldsymbol{x}_0,\boldsymbol{h})$
  END FOR
  概率融合与贪婪解码选取词元(如式7)：$h_t\leftarrow\mathrm{Fuse}(p_t^{(J)},\cdots,p_t^{(1)})$
  生成当前预测输出：$\boldsymbol{h}\leftarrow\boldsymbol{h}\oplus\boldsymbol{h}_t$
  IF $\boldsymbol{h}_t$== <EOS> THEN
    BREAK
  END IF
END FOR
初始化串联输入：$\boldsymbol{y}_0\leftarrow\boldsymbol{h}$
FOR $k=1,2,\cdots,K$ DO
  串联模型序列进行推理(如式5)：$\boldsymbol{y}_k\leftarrow M_k^s(\boldsymbol{y}_{k-1},\boldsymbol{p}_k)$
END FOR
RETURN $\boldsymbol{y}\leftarrow\boldsymbol{y}_K$

3.4 时间复杂度分析

为系统分析多模型协作结构在生成任务中的计算开销与推理效率，本文从理论层面对串联、并联及其组合协作结构的推理复杂度进行了统一建模与量化分析. 将每个模型在自回归生成中预测一个词元的过程抽象为单步计算开销函数$C(s)$，其中s表示模型的参数量或结构规模，用于单步推理的理论资源消耗. 在不考虑缓存复用、量化加速等工程优化的前提下，假设所有协作模型M_i具有相同规模与计算资源配置，即$C(s_i)=C(s)$，从而将复杂度分析聚焦于协作结构的组织方式与调度特性.

对于SI协作结构，模型按照固定顺序逐层推理，形成严格的链式依赖路径. 在每个生成时间步上，所有模型必须依次完成推理，因此整体时间复杂度为$C_{\mathrm{SI}}=L\cdot C(s)$，其中$L$表示串联模型的层级数. 该结构在逻辑建模与语义抽象方面具备优势，但由于推理路径无法并行，其时延随深度呈线性增长，整体吞吐效率受限.

PI协作结构则在每个生成时间步并行激活K个模型独立进行预测，并通过融合器聚合其概率分布生成共识结果. 尽管各模型理论上可并发执行，但由于融合器需等待所有模型完成预测后才能进行词元级合并，故整体推理时间受到最慢响应分支执行时间的限制. 根据“最慢路径主导”假设(假设2)，其时间复杂度为

$$C_{\mathrm{PI}}=\max_{i\in\{1,2,\cdots,K\}}C(s_i)+C_{\mathrm{agg}}\approx C(s)+C_{\mathrm{agg}}\tag{11}$$

其中,C_{agg}表示聚合过程的额外开销. 若不计融合开销,则简化为$C(s)$.

组合结构旨在融合串联和并联两类结构的优势,其推理路径由一个深度为L'的串联子结构与一个并行度为K'并联子结构共同构成. 整体时间复杂度可统一表示为

$$C_{\text{CI}}=L'\cdot C(s)+\max_{i\in\{1,2,\cdots,K'\}}C(s_i)+C_{\text{agg}}\approx(L'+1)\cdot C(s) \quad (12)$$

尽管S2P与P2S在形式上具有相同的理论复杂度表达,但二者在信息流形式与中间表示的语义特性方面存在显著差异. 结合两种策略的特点分析,S2P策略的串联阶段先对语义进行收敛性重构,使中间表示更稳定,随后进入并联阶段时模型间语义冲突显著降低,从而减少聚合开销与同步等待,因而预期具有更高的协作效率. 相反,P2S的前期并联阶段往往导致中间表示呈现较强的语义发散性和多路径冗余,使得后续串联阶段需完成更深层次的信息压缩与逻辑统一,从而可能间接增加整体推理时延.

4　实验设置

为了探讨SI、PI及CCI在多模型协作场景下的实际表现,本文设计了覆盖数学计算、知识理解与符号推理三类任务的系统实验. 不同于传统聚焦单模型能力的对比方式,本研究关注在存在性能差异与能力互补的多模型系统中,如何构建高效协作机制,以应对实际部署中模型能力不均衡、数据隔离与隐私限制等条件带来的性能瓶颈. 接下来,本文将分别介绍实验所选择的模型、任务数据集以及完整的实验方案设计.

(1)模型选择. 为了验证串联、并联及组合协作结构的有效性,本文选取了三个性能稳定、参数规模相近的国产轻量级大模型:Qwen-2.5-7B、DeepSeek-R1-7B和Yi-1.5-9B. 后续实验分别以Qwen、Ds和Yi作为其简称,以便于后续实验的讨论与分析. 相关模型参数配置如表1所示.

表1　实验模型配置与特性一览表

序号	模型名	代名词	参数量/B	发布单位
1	Qwen-2.5	Qwen	7	阿里巴巴
2	DeepSeek-R1	Ds	7	DeepSeek
3	Yi-1.5	Yi	9	零一万物

为模拟真实的分布式异构部署环境,三个模型分别运行于实验室服务器的三张NVIDIA A100(80 GB) GPU上. 该部署策略实现了物理资源的隔离,有效避免模型间的资源竞争与计算干扰,真实反映了多模型协作在异构设备上的通信调度与协同推理特性,为后续组合推理策略的评估提供了符合实际部署条件的实验基础.

(2)提示词策略. 如表2所示,为系统刻画不同协作结构中的提示词设计逻辑,本文根据协作路径中的模型所在的推理阶段进行划分. 对于单模型与PI结构因不涉及语义递进,仅设置统一的首阶段提示词以引导整体回答生成;SI结构包含三个连续模型节点,提示词依次承担"初稿—审校—验证"的链式任务角色;组合结构则根据路径结构分为两个阶段(如S2P的前序抽象与后序聚合,P2S的前期并行生成与后期逻辑整合),提示词相应引导模型完成语义提炼、多视角补充与一致性融合等子任务,从而确保各节点在协作链中履行明确职责.

表2　不同推理策略的提示词模板

协作策略	第一阶段模型	第二阶段模型	第三阶段模型
单模型/并联推理	你是一个专业的智能助手,请审慎思考并逐步推理,最后给出最准确的答案.问题:{question}.	—	—
串联推理	你是一个专业的智能助手,请审慎思考并逐步推理,最后给出最准确的答案.问题:{question}.	你是一位严谨的审校专家,请逐步审阅上一阶段的回答,纠正其中的错误与偏差,并补充遗漏的关键推理步骤,给出更准确的中间结论.助手回答:{result_1}.	你是一位经验丰富的验证专家,请对前述分析内容进行综合判断,凝练出最准确、清晰的最终答案.审校专家分析:{result_2}.
组合推理	你是一个专业的智能助手,请审慎思考并逐步推理,最后给出最准确的答案.问题:{question}.	你是一位经验丰富的验证专家,请对前述分析内容进行综合判断,凝练出最准确、清晰的最终答案.助手回答:{result_1}.	—

(3)数据集选择. 为了综合评估不同协作推理策略在多类型任务场景中的表现,本研究选取了三个具有代表性的数据集:GSM8K、CEVAL和Mathematics,分别覆盖数学计算、知识理解与符号推理三类典型任务,相关统计信息详见表3. 为确保实验条件一致性与结果的客观性,每个数据集均采用固定随机种子进行样本抽取,并从各种任务随机选取100道问题构建统一规模的测试集,为分析不同协作策略的性能差异与适用边界提供可靠的实验基础.

表3 实验数据集的基本信息与评估维度

序号	数据集	类型	测试能力	详细描述	示例
1	GSM8K	数学	数学推理与问题解决能力	包含约8 500道小学数学应用题,涵盖基础算术、代数、几何等知识点,要求模型具备多步推理能力.	问题:一个特殊的气球在水下每小时增加其前一体积的五分之二.如果它的原始体积是500 cm^3,在水下2小时后它的体积将是多少? 答案:980 cm^3
2	CEVAL	知识	综合知识与学科理解能力	多学科评估数据集,涵盖数学、物理、化学、生物、历史、地理等,要求模型具备综合知识理解能力.	问题:25 ℃时,将pH=2的强酸溶液与pH=13的强碱溶液混合,所得混合液的pH=11,则强酸溶液与强碱溶液的体积比是(忽略混合后溶液的体积变化)____. A. 11:1 B. 9:1 C.1:11 D.1:9 答案:B
3	Mathematics	符号	符号推理与规则推理能力	包含高等数学题目,如微积分、线性代数、概率论等,要求模型具备符号推理与公式推导能力.	问题:计算定积分:$\int_0^1 x^2\,dx$? 答案:1/3

(4)实验方案设计. 为了全面评估不同协作推理策略在多任务场景下的性能表现,本文构建了涵盖单模型推理、单一协作范式(SI/PI)与组合协作推理(S2P/P2S)在内的三类实验方案. 实验重点从推理准确率与推理时延两方面分析协作结构的性能表现,并进一步探讨协作结构设计对协同增益能力的影响机制. 具体实验方案如下.

(*a*)单模型推理. 以Qwen、Ds、Yi模型分别进行独立推理,统计其在各数据集下的平均准确率与平均推理时间,并作为多模型协作性能评估的基准参考. 此部分旨在揭示不同模型之间的推理能力差异,为后续分析协作结构带来的性能提升提供对照参考.

(*b*)单一协作推理. 构建了SI与PI两类基础协作结构. 对于每种结构,基于Qwen、Ds、Yi三个模型构建所有可能的全排列组合,共6组配置(例如Qwen→Ds→Yi、Ds→Qwen→Yi等). 该部分旨评估单一协作范式的性能差异,并分析不同协作路径下的性能波动与稳定性.

(*c*)组合协作推理. 设计了S2P和P2S两种组合策略. 与基础协作范式一致,组合结构同样基于Qwen、Ds、Yi构建六组组合配置,并在结构调度中保持协作层次的一致性. 例如S2P中先采用Qwen→Ds串联,再将其输出输入至Yi进行并联生成;P2S则由Qwen与Ds并联建模,最终由Yi串联接收并优化输出. 本方案重点探索组合协作结构能否有效整合不同协作范式的优势,从而验证在多任务场景下的整体优越性.

5 实验结果与分析

本节旨在系统评估多模型协作推理结构在不同任务维度下的性能表现. 通过构建单模型、SI、PI及组合结构(S2P、P2S)五种推理策略,并在三种典型任务(GSM8K、CEVAL、Mathematics)进行推理准确率与推理时延的系统对比,进一步探讨多模型协作的优势机制及应用潜力.

5.1 单模型、串并联与组合推理性能对比分析

如表4所示,本文对三种基础模型(Qwen、Yi、Ds)及其在SI、PI与组合(S2P、P2S)策略下的推理表现进行了对比分析. 实验以三个单模型的平均性能(即表中的"AVG"列)作为基线,分别从准确率与推理时延两个维度评估各协作模式的增益效果. 其中,符号"+"与"−"分别表示相较于基准的性能提升与下降幅度.

表4 单模型推理、串/并联推理与组合推理策略性能实验结果(准确率/推理时间)

任务类型	单模型推理				串/并联推理		组合协作推理	
	Qwen	Yi	Ds	AVG	SI	PI	S2P	P2S
GSM8K	67%/**19.29** s	26%/26.22 s	58%/21.47 s	50.33%/22.33 s	76%/33.17 s	74%/59.82 s	79%/24.81 s	**80**%/30.78 s
CEVAL	**74**%/**28.58** s	54%/42.30 s	29%/35.84 s	52.33%/35.57 s	67%/91.42 s	60%/51.72 s	68%/31.63 s	63%/37.50 s
Mathematics	86%/**20.97** s	50%/33.23 s	30%/34.86 s	55.33%/29.69 s	88%/45.01 s	74%/58.18 s	**91**%/22.31 s	**91**%/27.79 s
平均性能/%	75.67	43.30	39.00	52.67	77.00	69.33	**79.33**	78.00
平均时间/s	**22.95**	33.92	30.72	29.20	56.53(−27.33) ↓	56.57(−27.37)↓	26.25(+2.95) ↑	32.02(−2.82) ↓

注:加粗表示每一行中单模型与多模型协作推理结构中对应指标的最优值.

从图6和图7可以看出,尽管基础模型参数规模相近,但在多任务场景下性能表现存在显著差异. 其中,Qwen在三个任务中均表现领先,平均准确率达到75.67%. 相比之下,Yi与Ds整体表现相对较弱,分别为43.3%和39%. 进一步分析发现,Yi模型在CEVAL和Mathematics上的表现略优于Ds,而在GSM8K上则略显不足;Ds则在GSM8K上表现稍强,但在其他任务上整体准确率偏低. 这种分布不均的性能特征表明,不同模

型在任务维度上的能力偏差与互补性特征,由此也验证了构建多模型协作机制的动因:通过融合多个模型的优势输出,能够在多任务推理中实现性能互补与协同增益,从而缓解单模型在特定任务下的结构偏弱或知识覆盖受限等问题.

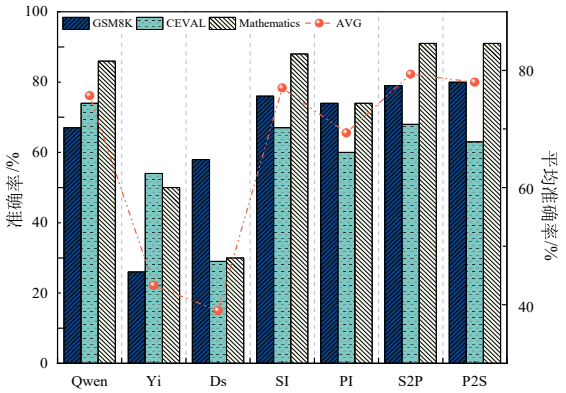


图6　单模型与协作推理策略在多任务下的准确率表现

在此基线上,本文进一步评估了不同协作结构在多任务场景下的推理表现. 结果显示,SI结构通过阶段性语义建模在一定程度上提升了整体推理能力,在三类任务上的平均准确率达到77%,相较于最强单模型提升了1.33个百分点,较平均性能基线提升了24.33个百分点. 而PI结构尽管具备并发解码与多视角建模能力,在缺乏一致性约束情况下,推理过程可能受弱模型干扰,导致准确率下降至69.33%,虽然相较于平均性能基线提升了16.66个百分点,但比最强单模型性能下降了6.34个百分点,表现出一定的不稳定性.

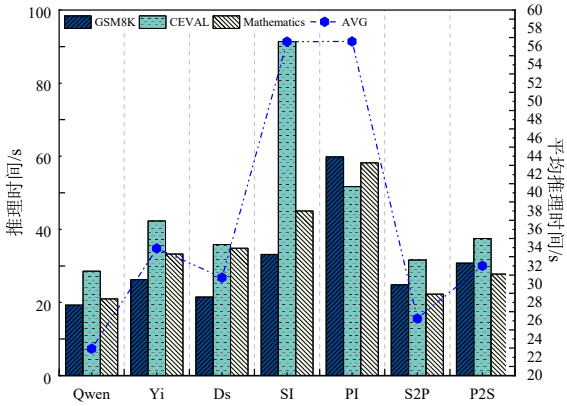


图7　单模型与协作推理策略在多任务下的推理时延表现

相比之下,组合协作结构在多任务推理中展现出更强的性能整合能力. 其中,S2P结构在所有单模型与协作结构中均取得最优表现,平均准确率达79.33%,相较最强单模型提升3.66个百分点,相较基线提升26.66个百分点;P2S结构同样稳定高效,平均准确率为78%,最高提升幅度达25.33个百分点,整体上优于SI与PI结构的表现. 该结果在一定程度上反映了组合结构集成串联与并联协作范式的互补优势,能够在逻辑收敛性与语义覆盖性之间实现较好的平衡,从而提升推理结果的一致性与泛化能力.

值得注意的是,在CEVAL单任务中,Qwen模型作为单模型表现最佳,其准确率甚至超越了所有协作策略. 这一结果表明,对于某些语义结构清晰、模型知识覆盖完备的特定任务场景中,协作策略未必带来额外收益,反而可能因融合误差或冗余信息引发性能扰动. 然而,从跨任务、跨领域的整体表现来看,组合协作策略总体表现更为稳健,显示出在多任务环境中整合模型能力、提升泛化与容错水平的潜力.

在推理时间方面,SI结构因模型间存在严格的顺序依赖,整体延时较高且波动明显,平均推理耗时达56.53 s,其中在CEVAL任务上最高可达91.42 s. PI结构尽管具备理论上的并发优势,但平均延时仍达56.57 s,且所有任务中均稳定维持在50 s以上,这表明其推理效率受限于内部结构性瓶颈. 为深入剖析该瓶颈的来源,本文对并联结构中单步令牌生成过程进行了细粒度性能分析(见表5). 结果显示,模型前向传播依旧主导了整体的延迟,而数据传输与词元聚合的开销均可忽略. 在三路并联结构下,各模型的单步前向推理存在显著时延差异,导致每一步的预测时间始终由最慢响应模型决定单步延时,形成一种动态的“最慢分支瓶颈”机制,成为并联结构响应速度的核心限制因素. 这一观察与本文提出的“最慢分支主导”假设(假设2)相符,并在实证层面补充说明了并联结构在同步机制下面临的效率约束.

表5　并联推理中前向计算、数据传输与结果融合耗时占比分析

耗时环节	均值/ms	标准差/ms	方差/ms	耗时占比/%
Qwen	60.38	7.07	50.00	—
Yi	96.53	10.07	101.39	—
Ds	105.32	7.84	61.39	—
最慢分支	106.13	9.71	94.32	99.9
数据传输	0.076	0.161	0.026	0.07
词元聚合	0.030	0.012	0.000 13	0.03

相比之下,S2P与P2S两种组合结构在推理效率上表现出更优表现,平均推理时间分别为26.25 s与32.02 s,均显著低于串联与并联结构. 其效率优势主要得益于结构设计的互补机制:S2P结构通过前期串联阶段实现语义聚焦与结构压缩,使得后期并联模型在统一且收敛的语境表示上进行独立解码,从而有效降低多语义感知过程中的冗余计算压力,避免了串联结构因从原始输入逐层构建逻辑链条所带来的全程线性延迟.

此外,组合结构在整体信息流路径上可视为一种“二级结构”,相较于传统三层串联模型减少了一次信息级联,在结构深度上进一步降低了时延开销. 然而,从实验结果看,组合结构的推理时延仍高于单个性能最佳的模型Qwen(22.95 s),这说明协作策略在提升性能的同时,不可避免地引入额外的协同开销. 因此,组合协作策略的优势并不在于追求单任务场景下的极致效率,而在于通过可接受的时延代价,换取在多任务场景下的准确率提升,从而实现更具泛化性的性能-效率的权衡.

5.2 模型路径对协作推理的影响分析

在多模型协作推理系统中,由于集成了多个异构基础模型,任务执行时通常存在多种模型路径可选. 然而,不同模型路径不仅影响信息传递的顺序,还可能显著影响最终协作推理的准确性和响应延迟. 为了系统性评估模型路径对协作性能的影响,本文基于SI、PI与组合(S2P、P2S)四类协作结构上,对三种基础模型(Qwen、Yi、DeepSeek)在不同组合路径下的协作性能进行全面实验与比较,旨在揭示协作结构中路径设计对协作效果的深层影响,为多模型编排与部署策略提供数据与理论支持.

5.2.1 串联与并联推理中的模型路径影响分析

为评估模型路径在SI与PI结构的影响,本文基于三个基础模型(Qwen、Yi、DeepSeek)构建了所有可能的全排列组合,以系统验证两类协作结构对模型顺序的敏感性. 在此基础上,分别在三类任务上量化分析不同路径配置下的推理准确率与执行时延,实验结果如表6所示.

表6 不同模型路径在SI与PI中的协作性能实验结果(准确率/推理时间)

任务类型	Qwen-Yi-Ds		Qwen-Ds-Yi		Yi-Qwen-Ds		Yi-Ds-Qwen		Ds-Qwen-Yi		Ds-Yi-Qwen	
	SI	PI	SI	PI	SI	PI	SI	PI	SI	PI	SI	PI
GSM8K/(%/s)	76/33.17	74/59.82	67/30.38	74/60.86	64/56.67	74/61.22	56/44.14	74/60.31	55/34.45	74/60.99	70/35.21	74/60.98
CEVAL/(%/s)	64/96.90	59/52.18	63/113.63	58/53.29	52/139.62	55/50.96	67/91.42	55/51.86	44/141.01	60/51.72	52/132.78	57/51.11
Mathematics/(%/s)	86/29.52	74/58.18	66/45.63	72/57.88	64/53.4	74/58	58/41.71	73/56.24	76/51.78	73/57	88/45.01	73/57.38
平均性能/%	**75.33**	**69.00**	65.33	68.00	60.00	67.67	60.33	67.33	58.33	69.00	70.00	68.00
平均时间/s	**53.20**	56.73	63.21	57.34	83.23	56.73	59.09	**56.14**	75.75	56.57	71.00	56.49

注:加粗表示SI与PI各路径在“平均性能”和“平均时间”维度上对应指标的最优值.

从图8~图10可观察到,SI结构在不同模型路径下的准确率表现出较强的敏感性. 其中,以Qwen为首位模型的路径(如Qwen→Yi→Ds)在三项任务中均取得最优或接近最优结果,平均准确率达到75.33%. 相对地,当Yi或Ds位于首位(如Ds→Qwen→Yi),准确率最低降至58.33%,性能波动区间高达17个百分点. 这在一定程度上说明,在串联结构中,前序模型对信息流具有关键的语义引导与特征压缩作用. 若弱模型位于首位,其生成的语义表示可能包含结构噪声或逻辑偏差,导致错误不断积累并放大,导致整体性能下降. 值得注意的是,当强模型Qwen位于末位时,仍可凭借其较强的知识表达与语义重构能力,在一定程度上弥补前序模型引入的偏差,保持相对稳定的输出效果,体现出一定的鲁棒性与纠错能力.

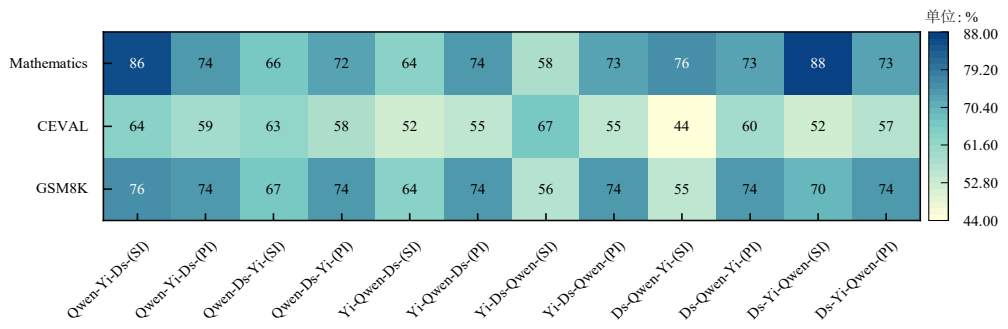


图8 SI与PI在不同模型路径下的协作推理准确性对比

相比之下,PI结构在模型路径的敏感性方面表现出明显减弱的趋势. 无论采用何种模型排列组合,在GSM8K任务中的准确率始终稳定在74%左右,在CEVAL与Mathematics任务中的性能波动幅度亦不超过±3个百分点. 整体来看,并联结构的平均准确率维持在68%,性能波动幅度始终控制在1个百分点以内. 这一结果表明,并联结构在架构层面具备较强的路径鲁棒性,能够在一定程度上降低系统对模型排列顺序的依赖程度. 值得一提的是,尽管从理论上并联结构应具有路径排列无关性,但为验证该假设在实际推理过程

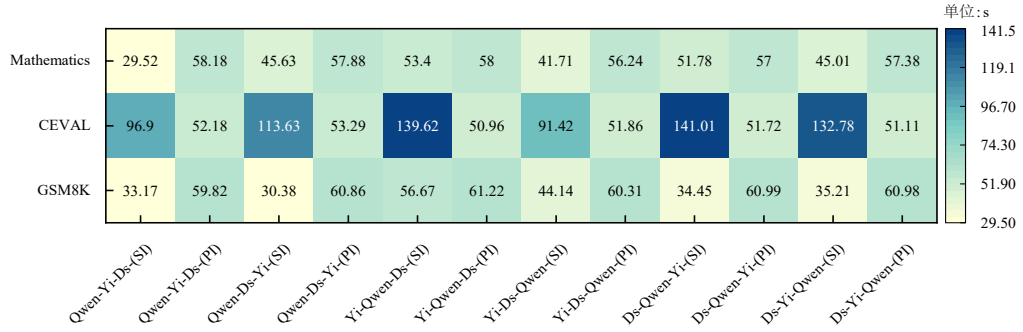


图 9　SI 与 PI 在不同模型路径下的协作推理延时对比

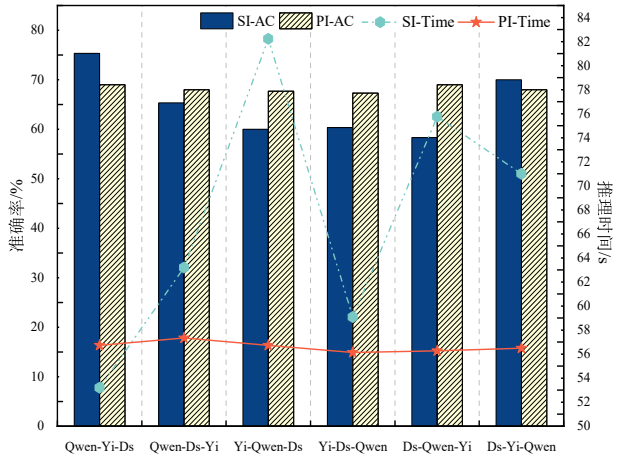


图 10　SI 与 PI 在不同模型路径下的平均协作性能

中的成立性,本文仍对其所有模型路径组合进行了系统测试,结果进一步说明并联结构在多任务场景下具有较强的稳定性.

在推理时间方面,如图 9 所示,SI 结构同样呈现出较强的路径敏感性. 以 CEVAL 任务为例,不同模型路径的推理时延最高达 141.01 s,较最低耗时相差近 50 s. 整体来看,串联结构在三个任务中的平均推理耗时波动最高可达 30.03 s,反映出一定程度上的不稳定性. 这一现象表现,路径设计不当可能导致前序模型传递的语义质量较低,不仅影响后续模型的判断准确性,也增加其语义修正负担,从而引入额外的推理延迟. 在 SI 结构中,这种语义质量波动与计算路径深度叠加的耦合效应,可能共同造成性能下降与时延扩大的双重影响. 相比之下,PI 结构在多任务中的时间波动极小,所有路径组合的平均推理时间均稳定在 56~58 s 之间,表明其延时主要由模型自身的计算耗时所决定,基本不受模型排列顺序的影响.

5.2.2　串并联组合推理中的模型路径影响分析

在前述实验基础上,本文进一步分析了组合协作策略中模型路径对组合协作推理的性能影响,重点分析了 S2P 与 P2S 两种结构在 3 类任务上的不同模型排列路径下的准确率与推理时延表现. 为此,本文对 3 种基础模型(Qwen、Yi、DeepSeek)在两类结构中的多种组合路径进行了系统测试与对比,实验结果如表 7 所示.

表 7　不同模型路径在 S2P 与 P2S 中的协作性能实验结果(准确率/推理时间)

任务	Qwen-[Yi-Ds]		Qwen-[Ds-Yi]		Yi-[Qwen-Ds]		Yi-[Ds-Qwen]		Ds-[Qwen-Yi]		Ds-[Yi-Qwen]	
类型	S2P	P2S	S2P	P2S	S2P	P2S	S2P	P2S	S2P	P2S	S2P	P2S
GSM8K/(%/s)	79/24.81	72/38.14	72/25.18	70/39.06	54/40.33	80/30.78	60/37.85	79/30.13	62/47.21	72/49.32	66/45.18	74/51.15
CEVAL/(%/s)	68/31.63	60/54.27	67/30.79	59/53.93	56/43.84	63/39.02	57/42.92	63/37.50	48/60.99	52/97.96	50/62.10	52/97.38
Mathematics/(%/s)	90/23.01	72/34.41	91/22.31	63/36.16	45/44.93	91/27.79	43/43.96	91/27.59	86/43.54	74/67.82	90/43.65	71/66.89
平均性能/%	**79.00**	68.00	76.67	64.00	51.67	**78.00**	53.33	77.67	65.33	66.00	68.67	65.67
平均时间/s	**26.48**	42.27	26.09	43.05	43.03	32.53	41.58	**31.74**	50.58	71.70	50.31	71.81

注:加粗表示 S2P 与 P2S 各路径在“平均性能”和“平均时间”维度上对应指标的最优值.

如图 11~图 13 所示,S2P 与 P2S 结构在不同模型路径下均展现出一定程度的性能波动,但其对路径顺序的敏感维度呈现出差异化特征. 具体而言,S2P 结构在准确率维度上的波段较大,路径间准确率分布范围为 51.67% ~ 79.00%,最大差值达 27.33 个百分点;而其推理时延波动相对较小,最大时间差值为 23.83 s. 相较之下,P2S 结构的准确率分布较为集中 64% ~ 78%,最大差值为 14 个百分点,但在推理时延上波动更为明显,路径间差值接近 40 s. 由此可见,S2P 更容易在准确率维度上受到路径配置的影响,而 P2S 的性能则在时间延迟维度上表现出更强的路径敏感性.

为进一步分析组合协作策略在路径敏感性方面对 SI 和 PI 结构的继承与缓解效应,本文对四类协作结构在不同路径下的准确率与推理时延波动进行了统计分析,结果如表 8 所示. 实验结果显示,S2P 虽在准确率上仍表现出一定的路径波动,但相较 SI 结构实现了平均

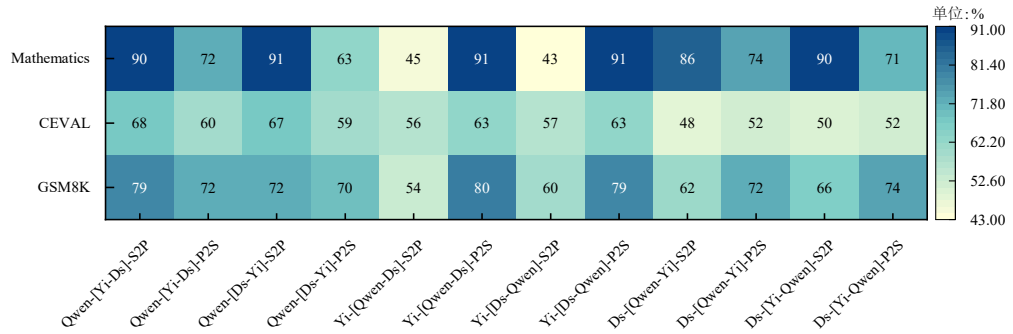


图 11 S2P 与 P2S 在不同模型路径下的协作推理准确性对比

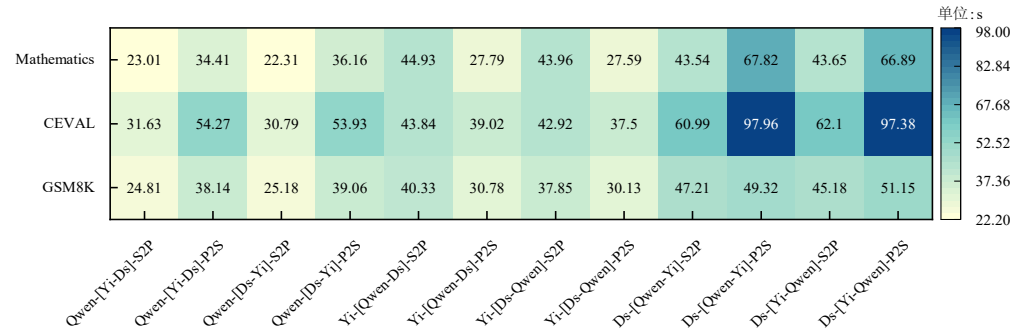


图 12 S2P 与 P2S 在不同模型路径下的协作推理延时对比

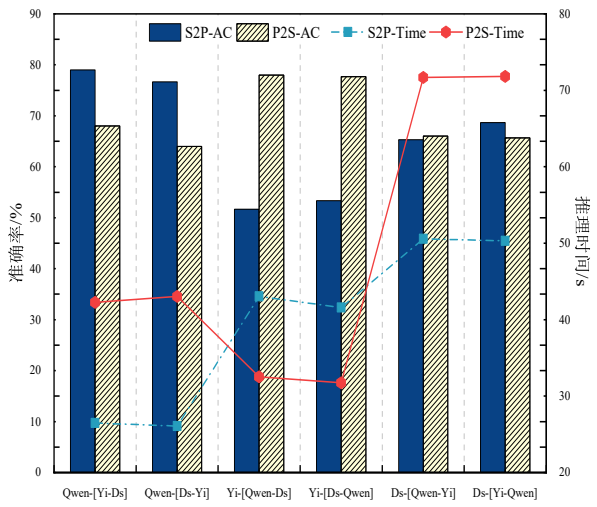


图 13 S2P 与 P2S 在不同模型路径下的平均协作性能

准确率的提升,并在整体推理时延与其波动性方面均有所缓解. 这一结果表明,S2P 在保留前序串联结构所具备的语义引导与建模能力的同时,通过引入并联子结构实现了对潜在偏差的语义补偿与纠偏. 尽管前序引导模型的性能差异可能带来准确率波动,但后续的并联阶段能够在聚焦的上下文条件下并行生成候选结果,从而降低同步等待成本,实现对推理延时的压缩与协作效率的提升.

相比之下,尽管 P2S 结构相较于 PI 结构引入了一定的路径敏感性,但其在不同路径下的准确率标准差

表 8 S2P 与 P2S 在路径层面的性能波动性指标

协作策略	平均准确率/%	平均时间/s	准确率(标准差)	推理延时(标准差)
SI	64.88	67.58	6.67	11.18
PI	68.17	56.67	0.69	0.39
S2P	65.78	39.68	11.45	11.00
P2S	69.89	48.85	6.28	18.36

低于 SI 与 S2P 结构,并在四种协作结构中取得最高的平均准确率,表明在准确率维度上对串联结构路径敏感性的影响具有一定的缓解作用. 这种稳定性可能得益于前序并联阶段对多源语义的融合与共识建构,能够在一定程度上抵消由引导模型性能差异带来的信息偏移,从而降低准确率的路径波动. 然而,在推理延时方面,P2S 结构呈现出更大的波动,且标准差亦高于 SI 结构,说明其在时间维度仍然保留了较强的路径敏感性. 分析认为,这一现象可能与前期并联阶段需生成更丰富语义上下文有关. 受“最慢分支主导”机制影响,推理过程存在不确定性与响应延迟积累;而后续串联阶段还需对这些冗余信息进行统一建模与逻辑重构,进一步拉大了路径间的响应差距.

值得注意的是,在组合协作结构中,仅调整并联子结构中的模型顺序对整体推理性能与时延影响相对较小,进一步验证了并联结构在路径配置方面所具有的稳定性. 由此可见,无论是“引导-补偿”型的 S2P 结构,还是“融合-决策”型的 P2S 结构,尽管在协作路径中不

可避免地引入了一定程度的路径敏感性,但整体上在多任务场景下均实现了准确率或推理时延维度的明显优化,展现出更为优越的性能与效率平衡能力.

5.3 协作推理结构在多任务中的性能分析

为深入探讨不同协作结构在多任务场景中的适应性差异,本文对SI、PI、S2P与P2S四种协作结构在GSM8K、CEVAL与Mathematics三类任务中的推理性能进行系统评估.具体分析了各结构在每项任务下的最优路径表现与全路径平均性能,并从准确率与时间效率两个维度进行综合对比,相关实验结果如表9与表10所示.

表9　不同任务上协作策略的推理准确率实验结果　　单位:%

任务类型	SI		PI		S2P		P2S	
	最优路径	平均性能	最优路径	平均性能	最优路径	平均性能	最优路径	平均性能
GSM8K	76	64.67	74	74.00	79	65.50	**80**	**74.50**
CEVAL	67	57.00	60	57.33	**68**	57.67	63	**58.17**
Mathematics	88	73.00	74	73.17	**91**	74.17	**91**	**77.00**

注:加粗表示每一行在"最优路径"与"平均性能"下的准确率最优值.

表10　不同任务上协作策略的推理时延实验结果　　单位:s

任务类型	SI		PI		S2P		P2S	
	最优路径	平均性能	最优路径	平均性能	最优路径	平均性能	最优路径	平均性能
GSM8K	33.17	39.00	59.82	60.70	**24.81**	**36.76**	30.78	39.76
CEVAL	91.42	119.23	51.72	51.85	**31.63**	**45.38**	37.50	63.34
Mathematics	45.01	44.51	58.18	57.45	**22.31**	**36.90**	27.79	43.44

注:加粗表示每一行在"最优路径"与"平均性能"下的时延最优值.

从图14与图15可以观察到,对于单一协作结构而言,SI结构在各类任务中最优路径的准确率相对较高,尤其是在CEVAL与Mathematics任务中分别达到67%和88%,均优于PI结构.然而,其平均准确率显著较低,反映出SI结构对模型路径具有较强的敏感性.该结果表明,若路径配置合理,SI结构能够充分发挥其在语义传导和深层次建模方面的优势,适用于结构化程度高、逻辑依赖明确的任务场景.相比之下,PI结构虽在最优路径上的准确率略逊一筹,但整体准确率更加稳定,3类任务中准确率均维持在较高水平.尤其在GSM8K任务中,其最优路径准确率达74%,已接近SI最优表现.这表明PI结构更适用于信息维度广泛、答案开放度高的任务,其并行结构有助于集成多视角语义,实现稳健而冗余的解码过程.

在组合协作策略中,S2P与P2S结构在3类任务中均在准确率与推理时延两个维度表现出优于单一SI与PI结构的综合性性能,兼顾了推理能力上限与稳定性输出,体现出良好的结构适应性与协同优势.具体而言,在CEVAL知识理解类任务中,S2P结构表现出相对更优的性能,其最优路径准确率达68%,且推理时间显著低于其他结构.该类任务涵盖多个子领域,具有较强的语义层级性与上下文依赖特征,对结构化建模与逻辑演绎能力提出较高要求.S2P结构通过"串联引导-并联补偿"机制,在前序阶段由串联模型聚焦全局语义压缩与主干建模,避免了多模型在语义分歧下反复迭代的共识成本.后序阶段则通过并联模型在统一语境下

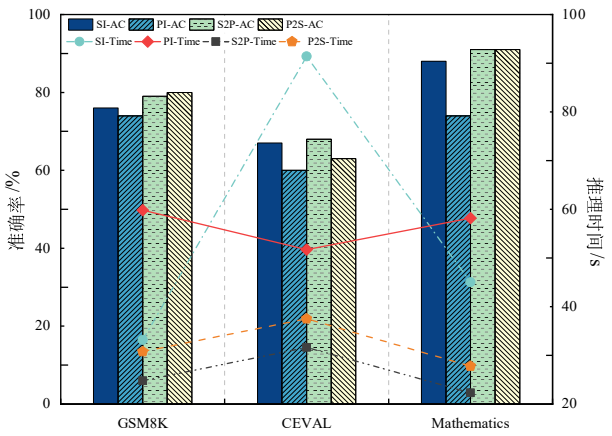


图14　不同协作推理结构在各任务下最优路径性能比较

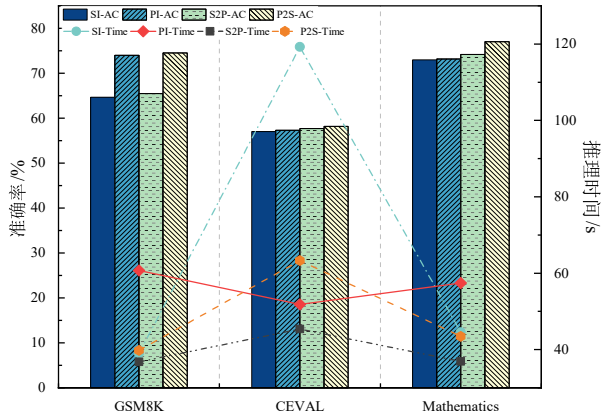


图15　不同协作推理结构在各任务下平均性能比较

进行多视角冗余补偿与结果修正,有助于提升语义一致性并增强输出的稳定性.

在GSM8K数学推理类任务中,P2S结构整体表现优异,其最优路径准确率高达80%,平均准确率为74.5%,显著优于S2P与其他基础结构. 该任务属于逻辑链主导型推理场景,问题求解步骤存在多种合理路径,答案空间存在较高的不确定性. P2S结构通过前序并联模块集成多模型的视角与认知偏好,能够有效提取多维数理特征并构建初步语义共识;后序串联模型则在此基础上完成链式整合与答案收敛,降低了误导路径带来的负面影响,从而提升了推理严谨性与鲁棒性.

在Mathematics符号推理任务中,S2P与P2S结构的最优路径准确率均达到91%,表现出相近的推理能力上限. 但在平均性能与时间效率方面表现存在差异:P2S平均准确率高于S2P推理,但S2P推理平均耗时仅为36.9 s,明显优于P2S结构. 这一对比表明,在结构化符号处理任务中,S2P的"结构主导+冗余补偿"机制有助于快速聚焦核心信息流、提高推理效率;而P2S则通过前段并联路径稳定提取特征、末端串联模型执行判断,适合对符号处理精度与抗干扰性要求较高的任务.

总的来说,无论是逻辑链长、路径分布广的GSM8K,还是知识覆盖广、语义结构清晰的CEVAL,或是对符号推理严谨性高的Mathematics,相较于单一协作结构,组合推理结构在性能上限与推理效率两个维度均表现出更为优越的整体能力. 这一结果说明,组合结构在协作范式层面有效融合了串联与并联的结构优势,构建出具有更强鲁棒性与计算效率的协同推理路径,能够在多任务场景下实现准确率与推理时延的双重优化.

6 结束语

本文围绕大模型在实际应用中面临的模型异构与推理协同问题,从模型协作层级与协作范式2个维度出发,系统提出了4种协作推理结构:SI、PI及其组合结构中的S2P与P2S. 为评估各结构在多任务场景下的适应性与性能差异,本文构建了覆盖3类典型任务(GSM8K、CEVAL、Mathematics)、3种基础模型(Qwen、Yi、DeepSeek)与24种路径组合的实验框架,从推理准确率与推理时延2个维度进行了系统对比分析. 实验结果表明,多模型协作结构整体上优于单模型推理,组合结构在多任务场景下进一步兼顾了准确率与推理效率,展现出更稳定的性能表现与更优的协同潜力. S2P结构在逻辑依赖性强、语义结构清晰的任务中更具优势,P2S结构则在处理解空间广泛、路径多样的任务中展现出更强的语义融合与抗干扰能力. 此外,实验还揭示了串联结构的路径敏感性与并联结构的效率瓶颈,为后续协作机制的优化提供了理论依据与实践参考.

未来研究可进一步探索更具自适应能力的协作推理机制,从任务特性出发动态调节串并联结构的组合方式,引入显式反馈与纠错机制以缓解路径敏感性影响,并在并联结构中设计更高效的聚合策略以突破最慢响应模型主导的结构瓶颈. 同时,在云边协同计算框架下,如何结合高效资源调度提升协作结构的运行效率,以及引入强化学习、自监督等机制实现任务驱动的协作优化,仍是值得深入探索的方向.

参考文献

[1] CHANG Y P, WANG X, WANG J D, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-45.

[2] WANG W H, CHEN Z, CHEN X K, et al. VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks[EB/OL]. (2023-05-25) [2025-09-20]. https://arXiv.org/abs/2305.11175.

[3] ZHANG D Z, YU Y H, DONG J H, et al. MM-LLMs: Recent advances in MultiModal large language models[C]// Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL, 2024: 12401-12430.

[4] 杨赟辉, 程虎, 魏敬和, 等. 面向Transformer模型边缘端部署的常用激活函数高精度轻量级量化推理方法[J]. 电子学报, 2024, 52(10): 3301-3311.
YANG Y H, CHENG H, WEI J H, et al. High-precision lightweight quantization inference method for prevalent activation functions in transformer models in edge device deployment[J]. Acta Electronica Sinica, 2024, 52(10): 3301-3311. (in Chinese)

[5] 徐刚, 刘志鹏, 冯骐, 等. 大语言模型在教育信息化中的实践: 规范、框架与应用[J]. 通信学报, 2024, 45(S2): 229-241.
XU G, LIU Z P, FENG Q, et al. Practical application of large language models in educational informatics: Specification, framework, and applications[J]. Journal on Communications, 2024, 45(S2): 229-241.

[6] 赖清楠, 金建栋, 周昌令. 基于大语言模型的网络威胁情报知识图谱构建技术研究[J]. 通信学报, 2024, 45(S2): 33-43.
LAI Q N, JIN J D, ZHOU C L. Research on knowledge graph construction technology for cyber threat intelligence based on large language models[J]. Journal on Communications, 2024, 45(S2): 33-43.

[7] QIU P C, WU C Y, ZHANG X M, et al. Towards building multilingual language model for medicine[J]. Nature Communications, 2024, 15(1): 8384.

[8] SUN Q S, YIN Z Y, LI X, et al. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration[EB/OL]. (2024-08-21)[2025-09-20]. https://arXiv.org/abs/2310.00280.

[9] JIN Z J, KLEIMAN-WEINER M, MIHALCEA R, et al. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents[EB/OL]. (2024-12-08)[2025-10-10]. https://arxiv.org/abs/2404.16698.

[10] EHTESHAM A, SINGH A, GUPTA G K, et al. A survey of agent interoperability protocols: Model context protocol (MCP), agent communication protocol (ACP), agent-to-agent protocol (A2A), and agent network protocol (ANP)[EB/OL]. (2025-05-23)[2025-06-10]. https://arXiv.org/abs/2505.02279.

[11] LI Q M, XIE Y. From glue-code to protocols: A critical analysis of A2A and MCP integration for scalable agent systems[EB/OL]. (2025-05-06)[2025-06-10]. https://arXiv.org/abs/2505.03864.

[12] CHEN Z Y, YANG X C, LIN J C, et al. Cascade speculative drafting for even faster LLM inference[EB/OL]. (2025-07-13)[2025-09-10]. https://arXiv.org/abs/2312.11462.

[13] JIN X S, REN X, PREOTIUC-PIETRO D, et al. Dataless knowledge fusion by merging weights of language models[EB/OL]. (2025-05-21)[2025-09-20]. https://arXiv.org/abs/2212.09849.

[14] YANG E N, WANG Z Y, SHEN L, et al. AdaMerging: Adaptive model merging for multi-task learning[EB/OL]. (2024-05-28)[2025-09-10]. https://arXiv.org/abs/2310.02575.

[15] SHNITZER T, OU A, SILVA M, et al. Large language model routing with benchmark datasets[EB/OL]. (2023-09-27)[2025-09-20]. https://arXiv.org/abs/2309.15789.

[16] FENG X C, HUANG Y C, LI B H, et al. Ensemble learning for heterogeneous large language models with deep parallel collaboration[C]//Proceedings of the 38th International Conference on Neural Information Processing Systems. New York: ACM, 2024: 119838-119860.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2023-08-02)[2025-09-20]. https://arXiv.org/abs/1706.03762.

[18] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 technical report[EB/OL]. (2024-03-04)[2025-06-10]. https://arXiv.org/abs/2303.08774.

[19] GRATTAFIORI A, DUBEY A, JAUHRI A, et al. The llama 3 herd of models[EB/OL]. (2024-11-23)[2025-06-10]. https://arXiv.org/abs/2407.21783.

[20] GUO D Y, YANG D J, ZHANG H W, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning[EB/OL]. (2025-01-22)[2025-06-10]. https://arXiv.org/abs/2501.12948.

[21] YANG A, YANG B S, ZHANG B C, et al. Qwen2.5 technical report[EB/OL]. (2025-01-03)[2025-06-10]. https://arXiv.org/abs/2412.15115.

[22] TEAM K, DU A G, GAO B F, et al. Kimi k1.5: Scaling reinforcement learning with LLMs[EB/OL]. (2025-06-03)[2025-09-20]. https://arXiv.org/abs/2501.12599.

[23] 张青龙, 韩锐, 刘驰. 云边协同大模型块粒度重训方法[J]. 电子学报, 2025, 53(2): 287-300.
ZHANG Q L, HAN R, LIU C. Cloud-edge collaborative retraining of foundation models at the block granularity[J]. Acta Electronica Sinica, 2025, 53(2): 287-300. (in Chinese)

[24] DONG X L, MOON S, XU Y E, et al. Towards next-generation intelligent assistants leveraging LLM techniques[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2023: 5792-5793.

[25] LEE M N, LIANG P, YANG Q. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities[C]//CHI Conference on Human Factors in Computing Systems. New York: ACM, 2022: 1-19.

[26] Medical large language model for diagnostic reasoning across specialties[J]. Nature Medicine, 2025, 31(3): 743-744.

[27] ZHU Y C, WU L, GUO Q, et al. Collaborative large language model for recommender systems[C]//Proceedings of the ACM Web Conference 2024. New York: ACM, 2024: 3162-3172.

[28] ZHU X, WANG Y, GAO H, et al. Recommender systems meet large language model agents: A survey[J]. Foundations and Trends in Privacy and Security, 2025, 7(4): 247-396.

[29] ZHU Y H, HE Z Y, HU H R, et al. MedAgentBoard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks[EB/OL]. (2025-10-30)[2025-11-10]. https://arXiv.org/abs/2505.12371.

[30] BREAZEAL C, CHAN Y, GHASSEMI M, et al. MDAgents: An adaptive collaboration of LLMs for medical decision-making[EB/OL]. (2024-10-30)[2025-09-20]. https://arxiv.org/abs/2404.15155.

[31] HAO Z X, JIANG H Q, JIANG S Q, et al. Hybrid SLM and LLM for edge-cloud collaborative inference[C]//Proceedings of the Workshop on Edge and Mobile Foundation Models. New York: ACM, 2024: 36-41.

[32] YANG Z M, YANG Y H, ZHAO C, et al. PerLLM: Person-

alized inference scheduling with edge-cloud collaboration for diverse LLM services[EB/OL]. (2025-05-23)[2025-06-10]. https://arXiv.org/abs/2405.14636.

[33] LI T L, LIU Q, PANG T Y, et al. Purifying large language models by ensembling a small language model[EB/OL]. (2024-02-19)[2025-06-10]. https://arXiv.org/abs/2402.14845.

[34] HOANG H, KHAYRALLAH H, JUNCZYS-DOWMUNT M. On-the-fly fusion of large language models and machine translation[C]//Findings of the Association for Computational Linguistics: NAACL 2024. Stroudsburg: ACL, 2024: 520-532.

[35] WAN F Q, HUANG X T, CAI D, et al. Knowledge fusion of large language models[EB/OL]. (2024-01-22) [2025-09-20]. https://arXiv.org/abs/2401.10491.

[36] BANSAL R, SAMANTA B, DALMIA S, et al. LLM augmented LLMs: Expanding capabilities through composition[EB/OL]. (2024-01-04)[2025-10-09]. https://arXiv.org/abs/2401.02412.

[37] VENKATRAMAN S, TRIPTO N I, LEE D. CollabStory: Multi-LLM collaborative story generation and authorship analysis[C]//Findings of the Association for Computational Linguistics: NAACL 2025. Stroudsburg: ACL, 2025: 3665-3679.

[38] NI A S, DESAI R, LI Y, et al. Collaborative reasoner: Self-improving social agents with synthetic conversations[EB/OL]. (2025-10-29) [2025-11-09]. https://openreview.net/forum?id=dye9w8IOV0.

[39] YANG S, LI Y F, LAM W, et al. Multi-LLM collaborative search for complex problem solving[EB/OL]. (2025-02-26)[2025-06-10]. https://arXiv.org/abs/2502.18873.

[40] 王建辉, 李哲涛, 伍涛, 等. Token级多模型并联协作推理[J]. 计算机学报, 2025, 48(11): 2579-2593.
WANG J H, LI Z T, WU T, et al. Token-level collaborative reasoning for parallel multi-models[J]. Chinese Journal of Computers, 2025, 48(11): 2579-2593. (in Chinese)

[41] YU Y C, KUO C C, YE Z Q, et al. Breaking the ceiling of the LLM community by treating token generation as a classification for ensembling[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg: ACL, 2024: 1826-1839.

[42] XU Y Y F, CHEN J H, WU J H, et al. Hit the sweet spot! Span-level ensemble for large language models[EB/OL]. (2024-09-27) [2025-09-20]. https://arXiv.org/abs/2409.18583.

[43] NIE L, DING Z M, HU E D, et al. Online cascade learning for efficient inference over streams[C]//Proceedings of the 41st International Conference on Machine Learning. New York: ACM, 2024: 38071-38090.

[44] NARASIMHAN H, JITKRITTUM W, RAWAT A S, et al. Faster cascades via speculative decoding[EB/OL]. (2024-10-21)[2025-10-10]. https://arXiv.org/abs/2405.19261.

[45] HU Z M, HUANG H. Accelerated speculative sampling based on tree monte carlo[C]//Proceedings of the 41st International Conference on Machine Learning (ICML). New York: ACM, 2024, 235: 19216-19251.

[46] XU H, YE J Y, LI Y T, et al. Can speculative sampling accelerate react without compromising reasoning quality?[C]//Proceedings of the 12th International Conference on Learning Representations (ICLR). New York: ACM, 2024 :1-7.

作者简介

刘忠仁 男,1996年1月出生于江西省赣州市. 现为暨南大学信息科学技术学院博士研究生. 主要研究方向人工智能与大语言模型推理系统.
E-mail: lzrisme@stu2024.jnu.edu.cn

王建辉 男,1997年12月出生于湖南省永州市. 现为暨南大学信息科学技术学院博士研究生. 主要研究方向为边缘计算和人工智能.
E-mail: tranfer98@foxmail.com

李哲涛 男,1980年1月出生于湖南省邵阳市. 现为暨南大学信息科学技术学院教授、博士生导师. 主要研究方向为云计算、智能网络、人工智能等.
E-mail: liztchina@hotmail.com

肖 勇 男,2000年10月出生于湖南省衡阳市. 现为暨南大学信息科学技术学院博士研究生. 主要研究方向为人工智能与数据隐私安全.
Email: xiaoyong@stu2022.jnu.edu.cn

曾曦玉　女,1999年4月出生于四川省宜宾市. 现为暨南大学信息科学技术学院博士研究生. 主要研究方向为人工智能、隐私安全及其应用.

E-mail: xyzeng@stu2025.jnu.edu.cn

李　俊　男,2003年6月出生于河南省信阳市. 现为暨南大学网络空间安全学院硕士研究生. 主要研究方向为大模型代理系统及其安全领域.

E-mail: koinu@qq.com

莫光峰　男,2002年2月出生于广东省茂名市. 现为暨南大学信息科学技术学院硕士研究生. 主要研究方向为人工智能与模型评估.

E-mail: moguangfeng2002@126.com