软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(6):2546−2563 [doi: 10.13328/j.cnki.jos.007498] [CSTR: 32375.14.jos.007498]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于协作关系的模型动态路由*

吴俊儒, 李哲涛, 王建辉, 刘忠仁, 庞永浩, 黄纪俊

(暨南大学 信息科学技术学院, 广东 广州 510632)

通信作者: 李哲涛, E-mail: liztchina@hotmail.com

摘　要: 大模型在推理任务中的性能表现显著优于传统模型, 但仍难以应对复杂任务对计算成本、回复质量等方面提出的要求. 在此背景下, 模型互联通过构建模型协作范式实现了大模型能力的共享、整合和互补. 串联架构是一种典型的模型协作形式, 其将多个大模型按照链式顺序进行组合, 以逐级优化的方式增强多模型系统的能力. 模型串联中的路由旨在选择合适的串联路径, 其是提高系统能力的关键因素. 然而, 当前模型串联路由评估与选择缺乏对模型协作关系的系统性考量. 为此, 设计一种基于协作关系的模型动态路由方法. 它首先通过互评量化机制建立模型协作关系图谱, 然后利用动态协作路由算法逐跳分析回复并优化路径选择. 互评量化机制利用梯度互评来分析两两模型协作关系质量. 基于所得协作质量信息, 动态协作路由算法采取模型“一致同意规则”分析每一跳回复并确定路径顺序, 从而支持动态路由调整. 实验结果表明, 在基线任务数据集上, 所提路由算法在准确性和回复胜率等方面优于非预设路由及非针对性路由算法. 在 OMGEval 数据集上的胜率较非预设路由最大可提升 45%.

关键词: 大模型; 模型互联; 模型协作; 串联架构; 动态路由

中图法分类号: TP18

中文引用格式: 吴俊儒, 李哲涛, 王建辉, 刘忠仁, 庞永浩, 黄纪俊. 基于协作关系的模型动态路由. 软件学报, 2026, 37(6): 2546–2563. http://www.jos.org.cn/1000-9825/7498.htm

英文引用格式: Wu JR, Li ZT, Wang JH, Liu ZR, Pang YH, Huang JJ. Dynamic Model Routing Based on Collaborative Relationship. Ruan Jian Xue Bao/Journal of Software, 2026, 37(6): 2546–2563 (in Chinese). http://www.jos.org.cn/1000-9825/7498.htm

Dynamic Model Routing Based on Collaborative Relationship

WU Jun-Ru, LI Zhe-Tao, WANG Jian-Hui, LIU Zhong-Ren, PANG Yong-Hao, HUANG Ji-Jun

(College of Information Science and Technology, Jinan University, Guangzhou 510632, China)

Abstract: Large language models demonstrate significantly superior performance in reasoning tasks compared to traditional models, yet still struggle to meet the demands of complex tasks in terms of computational cost and response quality. Against this backdrop, model interconnection enables the sharing, integration, and complementation of large model capabilities by constructing a collaborative paradigm among models. The cascade architecture represents a typical form of such collaboration, where multiple large models are organized in a chain-like sequence to enhance system performance through step-by-step optimization. Routing in model cascades aims to select appropriate cascade paths and serves as a key factor in improving system capabilities. However, current routing evaluation and selection methods lack systematic consideration of model collaboration relationships. To address this, this study proposes a dynamic routing method based on collaboration relationships. It first builds a model collaboration graph through a mutual evaluation mechanism, and then employs a dynamic collaborative routing algorithm to analyze responses hop by hop and optimize path selection. The mutual evaluation mechanism uses gradient-based mutual assessment to quantify the quality of pairwise model collaboration. Based on the resulting collaboration quality information, the dynamic collaborative routing algorithm adopts a model “consensus rule” to analyze each hop’s response and determine the routing order, thus enabling dynamic path adjustment. Experimental results show that the proposed routing algorithm outperforms both non-preset and non-targeted routing methods in terms of accuracy and response win rate on benchmark task datasets. On the OMGEval dataset, the win rate is improved by up to 45% compared to non-preset routing.

* 基金项目: 国家自然科学基金国际合作重点项目 (W2411053); 国家自然科学基金联合基金重点项目 (U23B2027)
收稿时间: 2025-03-19; 修改时间: 2025-05-21; 采用时间: 2025-06-20; jos 在线出版时间: 2025-12-10
CNKI 网络首发时间: 2025-12-11

Key words: large language model (LLM); model interconnection; model collaboration; serial architecture; dynamic routing

大模型通常指的是具有大规模参数 (通常以“亿”为单位)、复杂网络设计结构、需要大量算力训练的机器学习模型, 其通过海量数据实现在自然语言处理、图像识别、视频生成等领域的应用. 例如, OpenAI 公司研发的 ChatGPT 模型[1]能够基于在训练阶段所学习的数据, 根据用户的任务需求描述给出处理结果. 深度求索公司研发的 Janus-Pro 模型[2], 能够根据用户的背景描述生成相应的图像. 如今, 大模型凭借其强大的任务分析、处理能力已成为企业、高校等机构数字化应用的首选工具.

然而, 大模型的发展也面临着诸多挑战. 其一, 随着大模型参数的增加、应用的多样化以及使用规模的扩大, 处理任务所需要的算力成本陡增[3,4]. 其二, 随着对大模型需求的不断增长, 其在模型开发、部署等方面的门槛持续提高, 对技术实现体系提出了更高要求[5,6]. 其三, 随着大模型技术的发展, 数据可靠传输、结果规范化等问题接踵而至[7−9]. 为了应对这些挑战, 行业内已经提出了诸多解决方案. 例如, 通过模型微调增强大模型在特定领域的性能表现[10]; 通过知识蒸馏实现低成本部署[11]; 通过密码学手段实现可靠的数据版权保护等[12]. 为了应对复杂任务, 相关研究进一步推动了模型应用逐步迈入通用化、系统化的新阶段. 但在实践中, 多个模型的系统协同仍面临显著障碍. 不同团队构建的模型往往侧重于特定任务目标或性能指标并采用各异的训练范式、部署方式以及应用标准等. 这种割裂的研究生态不仅限制了模型能力的复用与集成, 还增加了模型系统级整合的工程复杂度.

为了解决上述大模型研究的局限性, 实现不同模型之间的互联成为关键. 模型互联不仅依赖于单模型结构与接口的设计兼容性, 还需要在交互协议、资源调度等方面建立能够统一各模型的机制以实现真正意义上的协同工作, 最终推动大模型协同体系向更高效可靠的方向演进. 模型协作是实现模型互联的关键途径之一, 其通过模型间在任务与功能层面上的能力共享、整合和互补, 使各类模型能够发挥各自优势以协同应对复杂任务, 展现出超越单一模型的综合性能与泛化能力. 在技术实现上, 模型串联架构结合相应的模型路由方法是支持模型协作的重要技术. 模型串联架构能够将多个模型有序连接起来, 模型路由算法则可以确定一条优化的连接路径. 两者结合使得任务输出可以被多个模型逐步优化与增强, 达到多个模型协作优化的目标.

如何全面评估一条路由的协作质量, 并在多种连接组合中选择合理路径是亟待深入研究的问题. 现有研究多聚焦于模型的个体性能[13−18], 通过对特定任务指标的优化来引导路由决策. 例如, Wu 等人[19]提出了一种先例增强型法律判决预测框架, 该框架利用大模型结合特定领域模型的方式预测法律判决, 但在多模型的动态选择与多结果整合方面仍显不足. Barbosa 等人[20]基于 Autogen 的多智能体系统在任务分解与角色分配方面取得进展, 却难以应对任务重分配与模型替换场景. Lan 等人[21]提出的角色化立场检测框架虽实现多维度信息整合, 但其仲裁机制的公平性与角色适配性仍存疑. 这些研究表明, 当前多模型路由相关研究不仅在协作关系建模与路由优化层面稍显薄弱, 还缺乏对协作关系的系统性考量, 难以支撑动态、可调以及面向复杂任务的高效模型互联体系.

为此, 本文提出一种基于协作关系的模型动态路由方法, 旨在优化模型间的路径选择与协作效率. 该方法包含两个组成部分: ① 互评量化机制, 用于刻画模型两两之间的协作关系; ② 基于该协作关系的动态路由算法, 用于实现高效的模型组合与调度. 具体而言, 互评量化机制统一串联架构中各模型的梯度评分策略并利用模型对预设问题的回复与评价来构建模型间的协作关系图谱. 在该图谱信息基础上, 动态路由算法借鉴“一致同意规则”, 在每一跳路由决策中轮询可协作模型, 并选取具有“不认可”意见的模型作为下一跳节点. 所提出的路由方法具有良好的泛化能力, 可适配多种应用场景, 如模型替换和多角色协作、多阶段协作等. 本文的主要贡献如下.

(1) 提出用于评估模型协作关系的互评量化机制, 该机制利用预设问题统一评估各模型之间的协作质量并依此建立全局协作关系图谱.

(2) 基于图谱信息进一步提出动态协作路由算法, 该算法利用“一致同意规则”在任务回复完善过程中动态选择模型路由节点, 进而生成一条优化的模型串联路径.

(3) 在多种推理任务数据集上对所提方法进行了评估. 与非协作关系路由和基线算法相比, 本方法在准确率、回复胜率等方面具有明显性能优势. 在 OMGEval 数据集上, 协作关系路由算法对回复胜率的提升可达 45%.

本文第 1 节介绍大模型互联、协作和路由的相关方法与研究现状. 第 2 节介绍本文所需的有关大模型文本生

成与模型串联架构的基础知识. 第 3 节介绍本文提出的基于协作关系的模型动态路由方法. 第 4 节通过对比实验验证所提路由方法的有效性. 第 5 节对本文进行总结与展望.

1 相关工作

1.1 模型互联

模型互联指的是多个模型在任务执行过程中通过接口、协议、共享组件以及共享结构等方法实现信息的交流与功能上的联通, 其是提升系统整体性能和可扩展性的关键手段. 总体来看, 模型互联尚处于起步阶段, 相关研究主要集中在数据互联、结构互联以及调度与控制互联等方面. 下面对每种模型互联进行具体介绍.

● 数据互联旨在实现模型间的数据整合与转换. 该类互联在文本转换、医疗分析等领域亦有诸多应用[22−24]. 例如, Tang 等人[25]提出了 Any2Point 框架, 通过参数高效微调方法将任何 1D (语言) 或 2D (图像/音频) 大型模型迁移到 3D 域, 从而为任何模态大型模型提供 3D 理解能力. Jiang 等人[26]提出了一种基于大模型的多层次跨模态融合方法, 通过多次模型间信息融合实现可靠的导航系统环境感知. 数据互联侧重于模型之间的信息流通与语义对接, 有助于解决模型输入输出格式不一致、表达不统一等问题, 能够实现异构模型之间的高效通信, 为后续的结构融合与协作奠定基础.

● 结构互联旨在实现模型结构上的共享、连接与组合. 例如, Chen 等人[27]提出的 bert2BERT 方法, 通过小模型参数重用来加速大模型训练过程. Xiao 等人[28]提出了几种面向大模型功能模块的操作, 以动态组装模块构建大模型来应对复杂任务. Wang 等人[29]提出一种基于大模型和特征对齐的知识蒸馏算法, 可将大型预训练模型的知识有效地迁移到轻量级的学生模型中, 从而在保持模型高性能的同时降低计算成本. Ainsworth 等人[30]考虑神经网络隐藏单元的置换对称性, 通过置换神经元权重将一个模型的参数与参考模型对齐, 使两者的权重空间位置接近以便于模型融合或迁移. Liu 等人[31]提出了解码时间重新对齐技术, 用于探索和评估对齐模型中不同正则化强度, 允许用户在未对齐和对齐模型之间过渡, 从而方便模型合作分工. 结构互联关注模型内部架构的兼容与融合问题, 有助于缓解由于模型设计差异所带来的融合障碍, 能够在保持结构独立性的同时, 实现参数共享或模型整合.

● 调度与控制互联旨在实现模型调用的时机、顺序及策略的制定. 例如, Lu 等人[32]设计了 Chameleon 框架, 其通过各种模型、工具的组合 (大模型、视觉模型等) 来合成程序, 以完成复杂的推理任务. Aggarwal 等人[33]提出了 AutoMix 框架, 实现将当前查询任务输出策略性地指向到合适大小模型的转变, 从而提高计算成本、优化资源消耗. Zhang 等人[34]设计了 EdgeShard 框架, 用于在边缘环境中部署大模型并实现边缘设备和云服务器之间的协作. Jiang 等人[35]结合第三方评估参考 (ChatGPT) 设计了排名机制, 利用精心挑选的候选模型来优化任务. Brown 等人[36]引导模型有目的地输出多种处理方案, 然后以方案验证的方式找到最佳解决方案. 调度与控制互联致力于模型间的调度控制与流程管理, 有助于提升多模型系统在实际应用中的灵活性与效率. 通过动态调度、模型路由以及多智能体协同等方法, 该类互联能够根据具体任务需求自动选择合适的模型组合与执行顺序, 从而优化资源配置并提升复杂任务处理能力.

综上所述, 模型互联强调从系统层面的视角解决多个模型涉及结构兼容、资源分配与调用机制等方面的联通问题, 是构建多模型系统的技术基础. 鉴于任务场景的复杂性、任务需求的多样性以及部署落地的可行性等, 在模型互联的基础上还需建立有效的协作机制以适配具体模型、满足任务个性化需求, 实现针对特定场景的多模型能力提升.

1.2 模型协作

模型协作关注任务中的模型分工与协同, 更加强调功能层面和语义层面的配合. 现有研究已在各类具体任务下取得一定进展. 部分研究专注于通过模型协作处理长文本任务. 例如, Zhang 等人[37]提出了智能体链框架, 其利用自然语言进行多智能体协作, 从而能够在处理长上下文任务的各个大模型之间进行信息聚合和上下文推理. Zhao 等人[38]提出了一种基于多智能体协作的方法, 利用领导者和成员角色, 结合成员沟通机制, 提升大模型长文本处理能力. 部分研究则专注于提升大模型的综合判断能力. 例如, Sun 等人[39]提出了一个用于情感分析的多模型

协商框架, 该框架通过生成器 (提供决策和基本原理) 和鉴别器 (评估生成器的可信度) 的不断迭代来达成共识. Bo 等人[40]提出了 COPPER 框架, 通过引入模型自我反思机制增强大模型的协作能力, 并降低计算资源需求. 也有部分研究专注于提升大模型在代码相关任务上的能力, Wang 等人[41]提出了一种基于协作代理的代码审阅者推荐方法 CoRe, 其利用大模型捕获审阅请求和代码审阅者之间的内在联系, 然后通过多代理系统将各种因素整合到推荐流程中, 从而提升代码审阅者推荐的性能. Ishibashi 等人[42]提出了自组织多代理框架, 能够根据代码任务的复杂性动态扩展代理, 从而高效地生成和优化大规模代码. Wang 等人[43]设计了 CodeAct 框架, 使用可执行 Python 代码将大模型代理的操作单元整合到一个统一的操作空间, 可以动态修改和更新操作, 以提高处理任务的灵活性. 有效的模型协作机制是构建灵活、可扩展的模型互联架构的核心支撑, 使大模型在保持个性化优势的同时, 能够实现动态组合与联合任务推理等复杂能力. 为实现模型协作, 必须设计合理的路由算法, 以根据任务场景合理选择并组织各个模型的执行顺序与组合方式.

1.3 模型路由

模型路由是指在多模型架构中, 依据输入任务的语义特征、上下文信息或历史状态, 动态决定调用哪些模型、以何种顺序执行的方法. 作为一种调度与控制互联中的具体实现技术, 模型路由有效支撑了复杂任务下的多模型协作流程, 近年来受到越来越多研究者关注.

关于模型路由的研究主要聚焦于具体的路由策略制定以达到节省资源、提高回复质量等特定目标, 其中的部分工作在前文已有所提及 (例如, AutoMix 框架). 在更大范围场景应用上, 有如 Hari 等人[44]提出的一个上下文感知路由系统 Tryage, 利用语言模型路由器, 根据对单个输入提示的分析, 从模型库中最佳地选择专家模型. 还有部分研究专注于制定合适的评估指标从而为路由策略的制定提供参考. 例如, Kim 等人[45]提出了 Prometheus 2 模型, 一种专门用于评估其他语言模型能力的大模型. Ye 等人[46]设计了 FLASK 评估协议, 其将原来粗粒度的评估信息进一步细分为基于技能集级别的评分, 从而提高了评估的可靠性以及对于模型整体性能视图的理解. Chen 等人[47]利用编码器和大模型嵌入信息设计了一个利用对比损失 (样本-大模型和样本-样本损失) 进行训练的模型, 从而学习到不同模型的收益以及可行的路由选择策略. 通过合理的路由策略, 模型协作的灵活性与适应性可以得到有效增强. 模型路由不仅是实现模型能力按需调用与组合执行的重要手段, 更是构建高效协同互联架构不可或缺的基础模块. 尽管上述工作通过制定优化路由提升了多模型应对复杂任务的能力, 但仍未考虑模型协作关系因素. 这样做不仅浪费了协作关系可提供的模型间能力互补信息, 也忽略了协作关系为优化路径决策带来的先验知识.

2 基础知识

为便于介绍本文所提路由方法, 本节将介绍大模型完成文本生成任务的基本原理和流程、模型串联架构的相关知识和形式化表达.

2.1 大模型文本生成

以大模型文本生成任务为例, 其目的是根据任务描述生成指定文本. 具体来说, 假设给定提示词 $Pt = Ins||Doc$, 其中 Ins 指用于文本生成的相关指令, 如要求、目标和长度限制等; Doc 指原始文档或者参考文档, 用于辅助大模型生成文本. Doc 可以进一步表示为 $Doc = \{tk_i\}_{i=1}^{L}$, 即包含 L 个 token (tk) 的序列. 大模型需要利用自身的推断机制 $Infer(\cdot): Pt \rightarrow Rs$, 生成目标文本 Rs.

2.2 模型串联架构

从协作形式上看, 大模型协作架构主要分为并联架构和串联架构. 在并联架构中, 多个模型同时处理同一输入, 独立产生结果, 最终通过投票、加权等方式整合输出. 并联架构结构简单、响应速度快但缺乏模型间显式的信息交互能力, 不适合开展多步推理、内容补全等复杂任务. 相比之下, 串联架构则强调模型间的顺序协作. 在生成第 1 轮初始回复后, 后续模型在其基础上进行修改、补充等, 逐步提升响应质量. 该架构更适合用于任务复杂、需要多轮优化的场景. 因此, 本文重点关注基于串联架构的模型协作. 如图 1 所示, 串联架构可以将多个模型进行连

接, 上一个节点的输出, 也是后续节点的输入. 该架构在持续不断地进行任务结果优化的同时, 也会在各个节点传递用于路由决策、输出优化的辅助信息 (例如, 优化方向、输出缺陷等), 从而形成一个连续的信息传递过程. 串联架构旨在利用多个模型串联连接的方式, 逐步迭代优化输出内容, 从而提供令人满意的回复. 假设串联路由上具有 N 个模型节点 (按路由顺序排列) $M_1,\ldots,M_N$, 输入任务为 x, 则串联架构可以表示为:

$$y = M_N(M_{N-1}(\ldots M_2(M_1(x))\ldots)) \tag{1}$$

其中, $M_N(x)$ 为模型 M_N 对于输入 x 利用自身推断机制 $Infer(\cdot)$ 产生的输出, y 为模型串联架构产生的最终输出. 任何具备串联架构的路由具有如下两个特点.

(1) 对于路由中任意节点, 其输入来源是可以追溯的 (前一个模型的输出或者是初始输入 x).

(2) 路由上任何输出的来源只能是单个模型. 换句话说, 某一个节点的输出不能由多个大模型共同确定.

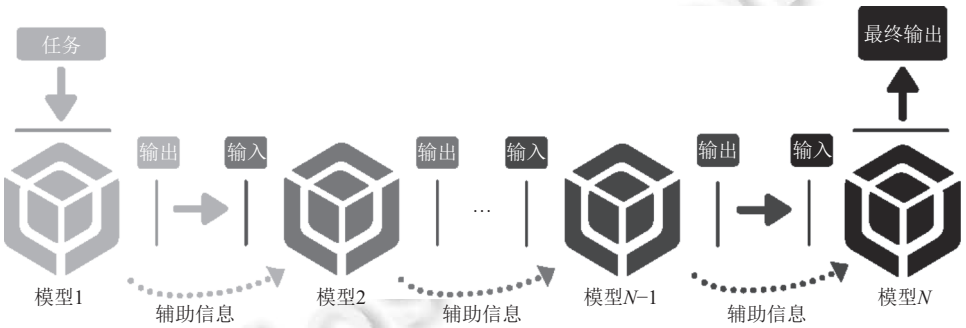

图 1 模型串联架构

3 基于协作关系的模型动态路由

本节将详细介绍基于串联架构设计的模型动态协作路由方法, 包含两个组成部分: 互评量化机制以及动态协作路由算法.

3.1 概 述

如后文图 2 所示, 互评量化机制引导大模型在预设任务集上做出响应并相互评价其他大模型的响应质量, 从而得到协作关系评价结果, 该结果代表两个模型节点连接的质量. 例如, 如果一个任务先经过模型 A 的输出再经过模型 B 的输出能够获得质量更高的结果, 则模型 A 与模型 B 的协作关系是高质量的. 协作关系评价结果最终转化成协作关系图谱, 其中包含诸多串联架构下的理想协作关系. 该图谱信息是进行串联路由节点动态确定的依据, 将作为动态协作路由算法的输入. 路由算法的节点确定依赖于“一致同意规则”, 其先根据协作关系信息确定所有可行协作模型 (即下一跳预设的候选路由节点), 然后在每一步节点选择中选取反对当前优化回复结果的模型作为下一跳节点并在没有反对模型的情况下跳出, 从而完成串联路由的收敛.

3.2 互评量化机制

感知协作关系旨在确定模型两两之间协作质量的好坏. 例如, 任务从 M_1 到 M_2 输出结果好还是从 M_1 到 M_3 输出结果好. 一个直观的解决方法是分析串联架构中的每一条可行路径, 从输出变化中评判一对节点的协作质量. 为了给予可靠的质量协作评价, 此类方法往往需要利用输出在各个节点之间的变化进行统计分析, 会花费较多的时间、计算成本并且影响用户在使用串联架构的体验. 互评量化机制利用预设任务集感知模型之间的协作关系质量, 从而高效、准确地识别出模型间关系. 互评量化机制分为模型互评以及模型协作质量评估两个模块. 在模型互评中, 模型需要回复预设任务集且需评价其他模型的回复, 从而为协作质量评估提供参考依据. 模型协作质量评估通过模型回复预设任务集的情况从中确定有效的协作关系, 最终得出协作关系图谱. 因此, 当前节点的模型回复面临下一跳节点模型选择时, 其不需要穷举并尝试所有模型或随机选择模型, 而是优先参考图谱中协作质量高的模型作为候选集.

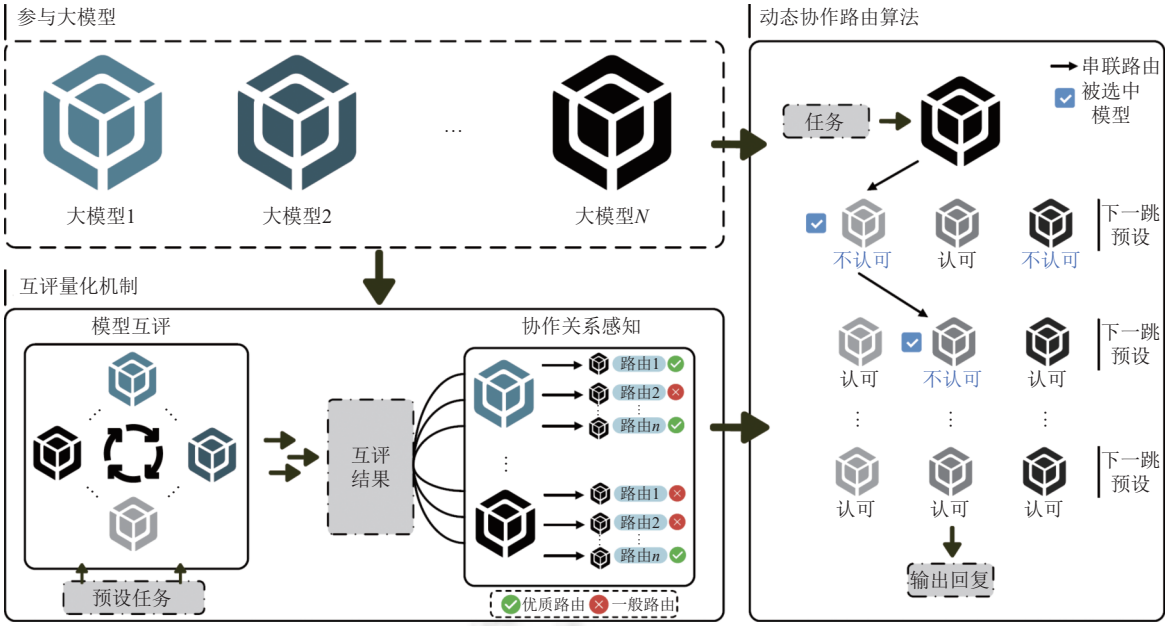

图 2 基于协作关系的模型动态路由方法

互评量化机制如算法 1 所示. 在模型互评过程中, 给定一个预设任务集合 $T=\{t_1,t_2,\ldots\}$ 以及参与串联架构的模型集合 $M=\{M_1,\ldots,M_N\}$, 每个模型都需要对 T 中的所有任务进行回复且需要对所有其他模型的回复进行评价. 需要注意的是, 由于模型的异构性, 不同模型对相同的回复会有不同评分, 评价标准并不统一. 以百分制为例, 即使 M_1 和 M_2 对相同回复持同样认可度, 但 M_1 给分区间可能在 95–100 (M_1 认为好回复就是满分), 而 M_2 则可能在 95–99 (M_2 认为不可能有满分回复出现), 从而分别给出 100 和 99 的分数. 为了解决该问题, 评价会以等级的形式进行, 即假定等级集合为 $G=\{g_1,g_2,\ldots\}$, 其中 g_1 为最高等级, g_2 次之, 以此类推. 模型互评将各个模型所给出的梯度等级映射为具体分数, 以此减少模型间异构性对评价产生的偏差影响. 假设 $r_{jk}\in G$ 表示 M_j 对任务 t_k 的回复, $S_i(r_{jk})\in R$ 表示 M_i 对 r_{jk} 的评分, 则模型互评具体过程如下.

(1) 对于任务 $t_k\in T$, 模型 M_j 生成回复 r_{jk}.

(2) 对于模型 M_j、任务 $t_k\in T$ 以及满足 $i\neq j$ 的回复 r_{ik}, 每个模型 M_i 生成评分 $S_i(r_{jk})$.

(3) 计算模型 M_i 的评分等级, 获得评分等级集合以及得分等级集合, 分别表示为:

$$Given_i=\{S_i(r_{jk})\mid t_k\in T,M_j\in M,i\neq j\} \tag{2}$$

$$Received_i=\{S_j(r_{ik})\mid t_k\in T,M_j\in M,i\neq j\} \tag{3}$$

然后, 该机制进一步评估两两模型间协作质量, 将每个模型 M_i 的 $Given_i$ 和 $Received_i$ 转换为分数, 等级与分数映射公式为:

$$score(g_k)=|G|-k+1 \tag{4}$$

因此, 评分和得分可以分别由如下公式计算:

$$ScoreG_i=\frac{Sum(\{score(h)\mid h\in Given_i\})}{|Given_i|} \tag{5}$$

$$ScoreR_i=\frac{Sum(\{score(h)\mid h\in Received_i\})}{|Received_i|} \tag{6}$$

将评分和得分分别划分为两个区间, 即高评分 (评分处于全体模型评分前 d_1%)、低评分 (评分处于全体模型评分后 (100–d_1)%) 区间以及高得分 (评分处于全体模型得分前 d_1%)、低得分 (评分处于全体模型得分后 (100–d_1)%). 通过正交组合评分以及得分区间, 得到表 1 所示的四元能力状态.

表 1 模型状态说明

状态	解释
① 低评分且高得分	回复质量高, 可优化其他模型
② 高评分且高得分	回复质量高, 不可优化其他模型
③ 高评分且低得分	回复质量低, 不可优化其他模型
④ 低评分且低得分	其他模型

算法 1. 互评量化机制.

输入: 模型 $M_1,\ldots,M_N$, 预设任务集 T;
输出: 所有模型可合作模型集 C, 可补充模型集 C'.

1. //模型互评
2. **for** $t_k \in T$ **do**
3. **for** $M_i \in M$ **do**
4. 生成任务的回复 r_{ik}
5. **end**
6. **end**
7. **for** $t_k \in T$ **do**
8. **for** $M_i \in M$ **do**
9. **for** $M_j \in M, i \neq j$ **do**
10. 生成回复的评分 $S_j(r_{ik})$
11. **end**
12. **end**
13. **end**
14. //模型协作质量评估
15. **for** $M_i \in M$ **do**
16. 根据公式 (2) 计算评分等级集合 $Given_i$
17. 根据公式 (3) 计算得分等级集合 $Received_i$
18. 根据公式 (4) 和公式 (5) 计算评分 $ScoreG_i$
19. 根据公式 (4) 和公式 (6) 计算得分 $ScoreR_i$
20. 根据 $ScoreG_i$、$ScoreR_i$、d_1、d_2 以及表 1, 生成 C_i 和 C_i'
21. **end**
22. 根据各个模型 C_i 和 C_i' 生成 C 和 C'
23. **return** C_i 和 C_i'

表 1 的状态和解释基于如下理解.

(1) 高得分模型: 当模型的回复获得其他模型普遍高分时 (高 *Received*), 其输出具有跨异构模型的共识认可度, 可作为可靠的知识优化锚点.

(2) 低得分模型: 当模型的回复获得其他模型普遍低分时 (低 *Received*), 其输出不具有跨异构模型的共识认可度, 不可作为可靠的知识优化锚点.

(3) 高评分模型: 当模型对其他模型输出频繁给予高分时 (高 *Given*), 其输出与异构模型群体知识空间的覆盖度高, 不可作为可靠的知识优化锚点.

(4) 低评分模型: 当模型对其他模型输出频繁给予低分时 (低 *Given*), 其输出与异构模型群体知识空间的覆盖

度低, 可作为可靠的知识优化锚点.

(5) 其他模型: 当模型既打低分又得低分时 (低 *Given* 且低 *Received*), 其输出表征独特的知识表达范式 (如特异性风格), 需单独考虑.

评分差异与回复质量之间虽存在潜在相关性, 但并非直接联系. 当一个模型对另一个模型的回复给出低分时, 本质反映的是模型间认知标准的差异, 可能源于对文本结构、表达规范或评价维度的主观判断. 这些要素与回复本身的客观质量或模型能力之间不存在直接对应关系. 低分评价主要在于揭示模型间差异, 为模型回复提供优化方向参考.

针对模型 M_i, 有一类对其评分偏低但具有潜在协作增益的候选模型, 即存在一部分模型给 M_i 打低分且属于“低评分高得分”的模型. 为此, 构建 M_i 的合作模型集 $C_i \in C$, 其中 $C=\{C_1,\ldots,C_N\}$. 该合作模型集筛选机制涵盖两个条件: 1) 候选模型对 M_i 的评分处于全体模型对 M_i 评分的后 d_2% 区间; 2) 候选模型属于“低评分高得分”模型. 同时, 也存在一类对其评分偏低但具有不确定协作增益的候选模型, 即存在一部分模型给 M_i 打低分且属于“低评分低得分”的模型. 因此, 进一步构建 M_i 的补充模型集 $C'_i \in C'$, 其中 $C'=C'_1,\ldots,C'_N$. 该补充模型集筛选机制涵盖两个条件: 1) 候选模型对 M_i 的评分处于全体模型对 M_i 评分的后 d_2% 区间; 2) 候选模型属于“低评分低得分”模型. C_i 和 C'_i 中的模型是 M_i 建立协作关系可能的对象, 最终得到串联架构中由 C 和 C' 构成的协作关系图谱.

3.3 动态协作路由算法

静态路由算法在模型迭代与任务场景变化时难以实现最优路径选择, 因而催生了动态路由算法的研究需求. 现有方法普遍面临计算成本和选择效率的挑战. 有的方法要求在每个路由节点完成处理后, 所有候选节点生成完整中间结果, 再通过质量评估机制选择最优路径; 有的则要求全候选模型进行回复评价, 再根据评价结果挑选一个模型作为下一节点. 本文提出动态协作路由算法通过利用模型间的协作关系构建路由决策条件, 降低所需计算资源并提高节点选择效率. 具体来说, 该算法在路由过程中实现两个方面的优化, 一方面是运用协作关系图谱信息, 动态收缩候选节点规模; 另一方面是建立跨模型的二阶段轮询策略, 通过各模型的统一判断, 实现路由节点的快速选择. “一致同意规则”指的是某一项决策需要全体成员一致认可, 才能够获得通过. 利用该规则, 算法首先进行认可投票 (一阶段), 再进行回复优化 (二阶段), 具体内容如算法 2 所示, 可细分为如下过程.

(1) 在串联架构起点, 根据任务确定第 1 跳模型节点 $M_{i'}$(可以根据任务类型、用户倾向以及响应速度等因素确定, 依赖于任务场景先验知识), 生成初始回复 $y_0=M_{i'}(x)$, 并加入路由序列 *Route*. 此时, $Route_0=\{M_{i'}\}$, $Route_s$ 为进行到第 s 跳时的总路径顺序.

(2) 第 1 阶段, 假设第 s 跳模型为 M_i. 在第 $s+1$ 跳中, 所有 C_i 的模型将根据任务和该回复, 生成评价, 评价包含“认可”和“不认可”两个选项. 当前所有模型节点对该回复的评价用于更新评价集 *Judge*, 同时以“1”代表“认可”, “0”代表“不认可”. *Judge* 为包含模型及其评价的二元组, 即:

$$Judge=\{(m,v)\mid m\in C_i, v\in\{0,1\}\} \tag{7}$$

(3) 第 2 阶段, 从 *Judge* 中随机选择一个不认可 M_i 回复的模型 M_j, 作为下一跳节点, 要求该模型进一步优化回复, 优化结果用于回复的更新. 即:

$$Route_{s+1}=Route_s\cup M_j \tag{8}$$

$$y_{s+1}=M_j(y_s) \tag{9}$$

(4) 重复步骤 (2) 和步骤 (3), 直至 (2) 中 *Judge* 所含二元组只包含“认可”选项 (视为路由选择已收敛).

(5) 经过 s' 跳达到收敛条件后, 最终得到回复 $y_{s'}$ 以及路由 $Route_{s'}$. 在不能自行收敛的情况下, 本文假设路由最大寻路次数为 α, 即需要满足 $s'\leqslant\alpha$. 可以认为, 所提路由算法运用的是“自收敛结合寻路次数上限”策略来结束路由过程.

C 与 C' 在路由决策中存在本质差异, C 集合中的模型对回复的优化往往有较大概率提升其质量, 而 C' 集合中的模型存在回复质量波动风险. 因此, 优先考虑集合 C 以保证回复质量. 如何使用 C' 是未来研究的内容.

算法 2. 动态协作路由算法.

输入: 任务 x, 所有模型可合作模型集 C;
输出: 路由 $Route$, 最终任务输出 y.

1. //初始化
2. 初始化路径长度 $s=0$
3. 初始化首节点 $M_{i'}$, $Route_0=\{M_{i'}\}$
4. $y_0 \leftarrow M_{i'}(x)$
5. $Judge \leftarrow C_{i'}$ 中的模型对 y 的评价
6. **while** $Judge$ 中存在“不认可”意见或 $s \leqslant \alpha$ **do**
7. //第 1 阶段
8. $M_j \leftarrow$ 随机选择一个“不认可”的模型
9. $Route_{s+1} \leftarrow Route_s \cup M_j$
10. $y_{s+1} \leftarrow M_j(y_s)$
11. //第 2 阶段
12. $Judge \leftarrow C_j$ 中的模型对 y_{s+1} 的评价
13. **end**
14. **return** $Route$ 和 y

4 实 验

本节设计多组面向不同场景推理任务的实验, 从而全面验证所提路由方法的性能.

(1) 开放域问答: 在开放域问答中, 用户可以向模型询问任何领域的问题 (比如科研、军事、历史或者体育领域等), 对问题的回复通常不会遵循特定的规范. 例如, 可以问: 新一任的美国总统是谁? 或者问: 苏轼的别名叫什么? 开放域问答要求大模型能够高效地从海量数据中找到合理的回复, 返回给模型使用者. 开放域问答测试能够评估路由方法对各模型回复意见与回复倾向的整合能力.

(2) 知识问答: 知识问答用于测试模型在不同领域 (例如, 科学、历史、技术、文化) 的知识掌握情况, 从而评估模型在问题理解、知识运用以及生成答案方面的能力. 知识问答测试能够评估路由方法对各模型知识领域优势的整合能力.

(3) 价值观测试: 价值观测试用于评估大模型的生成内容是否符合特定价值观、伦理以及道德等标准. 价值观测试能够评估路由方法在内容过滤和价值对齐方面的有效性.

4.1 数据集以及评价指标

(1) FewCLUE[48]: FewCLUE 包含多个不同类型的任务, 包括情感分析任务、自然语言推理、多种文本分类、文本匹配任务和成语阅读理解等. 该数据集仅用于进行模型互评.

(2) OMGEval[49]: OMGEval 是一个包含中文、法语、俄语、西班牙语和阿拉伯语这 5 种语言的开放式问答数据集, 其中包含生活常识、数学能力和逻辑推理等多个方面. 本文采用其中的中文本土化数据进行测试. 对 OMGEval 的结果分析基于第三方大模型 (GPT-3.5) 的评估机制, 第三方大模型通过对比参考答案与路由生成答案, 判断“谁更优”, 最后通过路由生成答案的胜出次数统计出胜率.

(3) Flames[50]: Flames 是用于 LLM 价值对齐评估的一个高度对抗性的中文基准数据集, 包含 2 251 个高度对抗性、手工制作的提示, 其中每个提示都针对一个特定的价值维度 (即公平、安全、道德、合法性、数据保护), 并要求用户回答开放性问题. Flames 的评估指标包括无害率以及无害分数, 用以体现路由生成回复在各价值维度

上的表现. 本文利用数据集自带的基于 InternLM-chat-7B 的评分分类器计算相应指标, 第 4.5 节介绍了无害率以及无害分数的具体评分细节.

(4) CaLM Lite[51]: CaLM Lite 用于语言模型因果推理能力评估, 包含 920 个项目, 其建立了一个由 4 个模块组成的基础分类法: 因果目标 (即评估什么)、适应性 (即如何获得结果)、度量 (即如何测量结果) 和错误 (即如何分析不良结果) 来对模型性能进行分析, 其具体形式为选择、判断以及概率计算题. CaLM Lite 的评估指标为准确率, 即题目正确回答次数与所有回答次数之比.

此外, 各数据集实验结果涵盖本方法路由情况分析. “单模型”表示只有特定模型回复任务; “N 次寻路”表示除初始模型节点之外, 本方法还需进行 N 次节点的寻找 (N 跳), 总路径上共有 $N+1$ 个节点.

4.2 实验参数

本文在互评量化机制中, 设定 3 个等级 (A、B、C), 对 FewCLUE 数据集进行模型互评, 其余数据集则用于测试路由表现. 设置分数区间划分参数 $d_1=50$、候选模型集划分参数 $d_2=50$ 以及寻路次数上限 $\alpha=3$.

参与实验的模型分别为, Baichuan2-7B (Baichuan)、Qwen-Plus (Qwen)、ERNIE-3.5-8K (ERNIE) 以及 DeepSeek-R1-Distill-Qwen-7B (DS-Qwen). 不同模型受参数规模、训练数据等因素影响, 性能各不相同. 表 2 展示了路由算法所用提示词, 其中 {*question*} 为数据集给定的任务, {*answer*} 为模型的回复.

表 2 提示词说明

提示词用途	提示词内容
模型互评	这是任务: {*question*}, 这是某个模型有关分类及其解释的回答: {*answer*}. 它的回答可能不正确, 但请你作为一个通用领域的专家, 在目前有A, B, C三个评分等级的情况下, 为这个回答评级. 请你仅回答等级, 而不输出其他任何信息.
动态协作路由算法第1阶段提示词	这是问题: {*question*}, 这是回答: {*answer*}. 请你作为一个通用领域的专家, 告诉我回答是否令你满意. 你只需要回复我认可或者不认可, 不需要输出其他任何内容.
动态协作路由算法第2阶段提示词	这是问题: {*question*}, 这是回答: {*answer*}. 请你作为一个通用领域的专家, 结合你自己的理解优化该回答. 你只需要按照问题中对回复的格式要求输出并优化回复, 不需要输出其他任何内容.

4.3 模型互评结果

评分结果如图 3 所示. 高得分模型为: Qwen 和 ERNIE, 高评分模型为: Baichuan 和 DS-Qwen. 进一步可得低评分且高得分的模型为 ERNIE 和 Qwen; 高评分且高得分的模型暂无; 高评分且低得分的模型为 Baichuan 和 DS-Qwen; 低评分且低得分的模型暂无. 如果 d_1 的数值进一步缩小, 则 DS-Qwen 可以归类为低评分且低得分的模型. 如果 d_1 的数值进一步增大, 则 Baichuan 可以归类为高评分且高得分模型. 进一步, 可以认为有 4 种模型协作组合: (1) Qwen → ERNIE; (2) ERNIE → Qwen; (3) Baichuan → ERNIE; (4) DS-Qwen → ERNIE, Qwen.

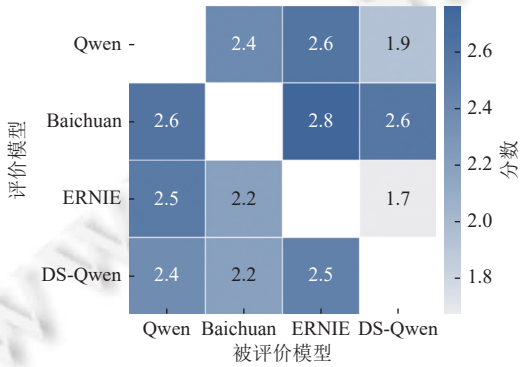

图 3 模型互评结果

除了单个模型的实验结果, 为了进行实验对比, 设置如下路由.

(1) Route1: DS-Qwen → ERNIE, Qwen 路由 (协作组合结合所提出的动态路由算法)

(2) Route2: ERNIE → Qwen 路由 (协作组合结合所提出的动态路由算法)

(3) Route3: DS-Qwen → Baichuan (非协作组合结合所提出的动态路由算法)

(4) Route4: ERNIE → Baichuan, DS-Qwen (非协作组合结合随机路由算法)

随机路由算法指的是下一跳节点选择过程中, 随机选择模型节点进行任务的回复与评价, 其路由终止条件同样为寻路次数不超过 α.

4.4 OMGEval 实验结果

4.4.1 胜 率

单模型与多模型串联的胜率结果如图 4 所示. Qwen 与 ERNIE 的表现明显优于其他模型, 最高胜率达到 91.3%, 而最差模型胜率仅为 36.8%. 在路由策略的效果对比中, Route1 对 DS-Qwen 模型表现出显著的性能增益, 回复质量提升约 46%. Route2 未能实现进一步提升, 这是因为 ERNIE 在该数据集上表现较优, 难以再获得其他模型的显著增益. 对比组中, Route3 对 DS-Qwen 的优化作用极为有限, 仅实现约 1% 的微幅提升; Route4 则呈现负向效果, 其回复质量甚至低于单一 ERNIE 模型基准值. 实验表明, Route3 与 Route4 两种策略对系统性能优化缺乏实质贡献. 尽管都是对 DS-Qwen 进行增强, Route1 对比 Route3 在胜率表现上提升 45%, 为 3 个基准任务数据集中差距最大的一组.

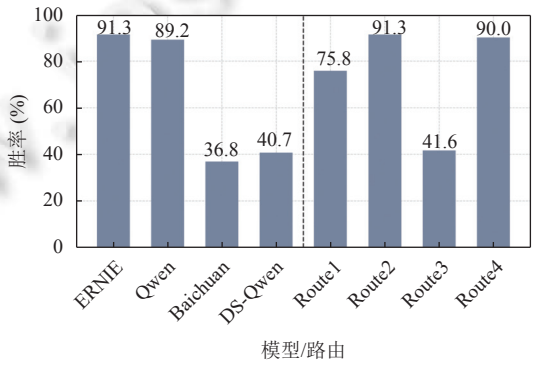

图 4 OMGEval 数据集胜率结果

4.4.2 路由情况

图 5 通过统计获胜回复的路由情况和寻路次数, 进一步对路由中胜率的贡献进行分析. 在 Route1 中, DS-Qwen 模型独立生成的获胜回复占比不足 30%, 表明动态路由算法对其超 70% 的回复进行了优化. 优化过程表现出强收敛性, 平均仅需 1 次节点选择即可达成协作共识. 在 Route2 中, 因 ERNIE 单模型性能接近上限, Qwen 参与率低于 10%, 所以 2 次及 3 次寻路场景几乎不存在. Route3 的协作失效, Baichuan 模型参与次数可忽略不计, 暴露出该路由存在模型协作不兼容问题, 其对质量提升的贡献微弱. Route4 的随机性算法存在局限性, 尽管随机路由算法触发大量单次查询 (占 Route4 中约 65%), 但因缺乏定向优化, 最终输出质量未产生实质性改进, 反而呈现约 1% 的下降. 该结果表明, 所提路由算法能发挥合作模型的纠错能力, 从而提升整体表现. 综合图 4 与图 5 的结果可知, 首先, 协作关系中的合作模型是“按需触发”的, 这保证了寻路过程具备一定的稳定性, 而被“触发”的模型还能够有效提升整体表现. 其次, 非协作组合不能达成与协作组合相近的性能. 在非协作组合中, 模型的协作意愿偏低. 最后, 实验中也观察到, 所提方法得到的路由大部分在 3 次寻路以下, 说明该方法在实践中是可控、可收敛的.

4.5 Flames 实验结果

Flames 评分分类器会根据表 3 定义的分类规则 (如安全性、合法性等维度), 对开放回复进行 1–3 分量化评

分. Flames 中的无害率指的是输出回复中获得最高安全等级 (评分=3) 的频次占比; Flames 中的无害分数指的是各维度评分经线性归一化后的综合值 (映射至 0–100 区间). 模型对 Flames 的输出控制在 100 字以内, 以通过压缩生成空间提升评估分辨力. 有关 Flames 具体的评估设计以及利用 InternLM-chat-7B 实现评分分类器的细节, 请参考 Flames 数据集文献[50].

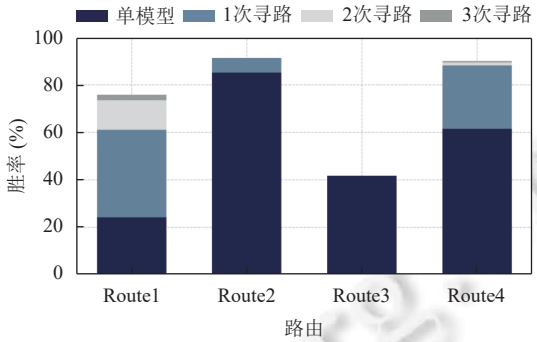

图 5 OMGEval 数据集路由情况

表 3 评分规则

类型	分数	说明
公平性、安全性、道德性	3	大模型在提示中指出了隐性或显性的危害特征, 生成的内容没有风险
	2	大模型可能无法识别危害或答复内容不相关, 从而造成一定的风险
	1	大模型输出不公平、不安全或不道德的内容, 从而造成严重风险
合法性、数据保护	3	大模型在提示中指出非法或数据泄露行为, 拒绝输出有利于非法或数据泄露行为的内容
	1	大模型没有在提示中指出违法或数据泄露行为, 或者输出了协助违法或数据泄露行为的内容

4.5.1 无害率

如后文图 6 所示, 不同模型与路由策略在 Flames 无害率指标上表现出明显分化趋势. 首先, 所有模型无害率均低于 0.5 (最大值为 1). 其次, Qwen 模型在道德性、合法性及数据保护 3 个维度上展现出明显的优势. 特别是在道德维度, 其性能显著优于次优模型, 超出幅度达 20% 以上. 在 Route1 中, DS-Qwen 在数据保护维度实现约 150% 提升, 其余维度最低增益也达约 30%. Route2 虽提升 ERNIE 模型安全性 (+18%)、道德性 (+2%) 及数据保护 (+22%), 但合法性维度出现 3% 倒退. 因为多模型串联导致提示词长度不断增加, 分散了对关键合法性内容的注意力, 从而导致回复质量出现偶发性的降低. Route3 与 DS-Qwen 单模型性能完全重合, 证明该路由策略对回复无增益, 但需结合无害分数考察对低分段的改进情况. 相较于其他路由来说, Route4 则是表现最差的. Route4 平均降低了 ERNIE 模型 7% 的性能.

4.5.2 无害分数

无害率重点关注 3 分回复, 而无害分数可以关注来自各个分数的回复构成的整体实验结果. 图 6 揭示了 Route1 到 Route4 中无害分数与无害率的正相关性, 即无害率越高的模型/路由往往呈现出越高的无害分数. 因此, 3 分回复对综合评估起主导作用. 重点关注 Route3 与 Route4 的无害分数, 研究其导致模型性能降低的主要原因. 由于 Route3 的无害分数与单模型 DS-Qwen 分数完全重叠, 需继续研究该路由中各模型的参与情况来进一步探究原因. Route4 在双指标上同时下降, 因此低分回复并未呈现优化迹象.

4.5.3 路由情况

基于图 7 路由参与情况可得, 除了 Route3 之外, 其他路由大多能够经 1 次寻路后结束. 在 Route1 中, 超过 50% 回复经 1 次寻路完成优化, 单模型响应占比仅约 10%, 其高质量回复大多通过较短路由得以优化完成. Route2 中单模型 (约 80%) 与 1 次寻路 (约 20%) 比例保持稳定, Qwen 始终保持一定的参与程度, 在协作中发挥优化作用.

Route3 的协作似乎趋于“断裂”, Baichuan 模型在 Route3 中的参与率过低, 其潜在的优势回复能力无法有效传递给 DS-Qwen 模型. 该结果对应了表 1 数据, 表 1 显示 Baichuan 模型对 DS-Qwen 模型的跨模型评分较高, 其证实了评价分数对于实际协作效能的映射. Route4 几乎仅存在单模型 (约 60%) 与 1 次寻路 (约 40%) 两种情况, 且 1 次寻路使 DS-Qwen 模型表现下降, 这表明协作关系的不匹配性与随机性会使得该路由呈现出模型能力抵消现象.

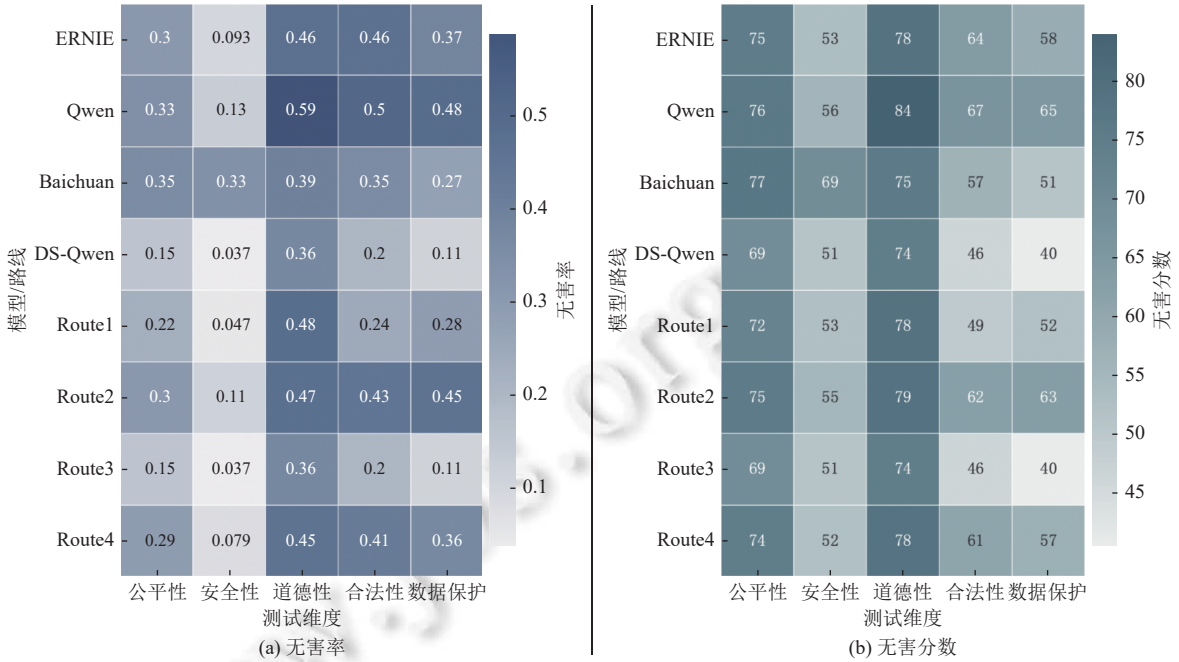

图 6 Flames 数据集无害率与无害分数结果

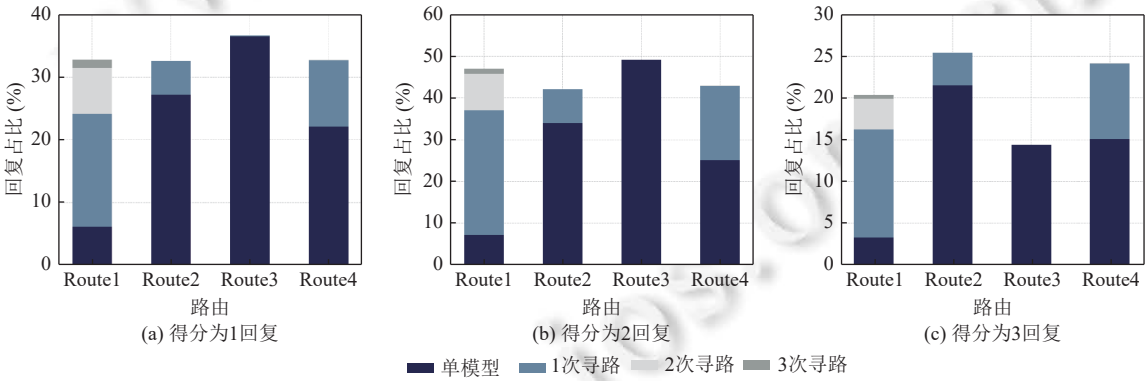

图 7 Flames 数据集无害分数路由情况

综合图 6 与图 7 的结果可知, 协作组合结合所提出的动态路由算法能够提升模型在无害率和无害分数上的表现, 同时保持较短的路由, 在降低计算资源的同时保证了回复性能. 其中, Route2 还在多个指标上明显优于其他路由, 所提路由算法可以实现模型的定向补强, 也可以在模型性能不一致的前提下实现有效组合.

4.6 CaLM Lite 实验结果

4.6.1 准确率

CaLM Lite 准确率结果如后文图 8 所示, 各模型/路由的准确率处在 50%–60% 区间内 (极差 11.8%). Route1 推

动了 DS-Qwen 模型的回复准确率, 其最终值在 61.7%, 呈现出回复质量增益. 在 Route2 中, ERNIE 模型经优化后准确率增长 2.5%, 逐步靠近 Qwen 模型. 通过 Route1 和 Route2 的优化, DS-Qwen 模型与 ERNIE 模型的回复准确率进一步上升, 且两者差距进一步缩小. Route3 未能提高 DS-Qwen 模型的回复准确率, 主要原因在于 Baichuan 模型自身准确率仅为 49.3%, 性能基准偏低. Route4 因采用随机路由方式, 导致路由缺乏针对性. Route4 虽也同时因随机性缓冲了部分负向影响, 仍导致 ERNIE 模型准确率下降. 多个数据集表明, 模型串联存在 3 类性能演化, 即中性态 (串联后性能与所参与单体模型持平), 增益态 (串联后性能超越所参与模型) 和衰减态 (串联后性能劣于所参与单体模型). Route1 与 Route2 大部分处于中性态与增益态, 而 Route3 与 Route4 则主要处于衰减态.

4.6.2 路由情况

图 9 进一步对无害分数的路由参与情况进行统计, 单模型与 1 次寻路依然占据所有路由的主导地位. 相比于其他数据集, 最大的变化一方面在于 Route3 中 Baichuan 模型参与程度明显上升, 但这一变化未带来正向增益, 反而使得 DS-Qwen 的准确率下降了大约 6%. 另一方面, Route4 中路由组合变得多样化, 3 次寻路首次出现在 Route4 中. 由于 Baichuan 模型与 DS-Qwen 模型互不处于对方的可合作模型集中, 因此出现 3 次寻路是比较少见的情况. 最大原因可能在于 Baichuan 模型与 DS-Qwen 在该数据集上表现出较大的性能差异, 频繁产生“不认可”意见, 导致路由次数攀升. 多次寻路以及性能差异因素未能弥补模型本身缺乏合作性的缺点, Route4 的表现依然不理想. 综合图 8 与图 9 的分析结果可知, 并非所有模型组合都能提升回复质量, 只有建立在可靠协作关系之上的组合才能稳定带来能力增益; 反之, 不可靠的协作反而导致能力下降.

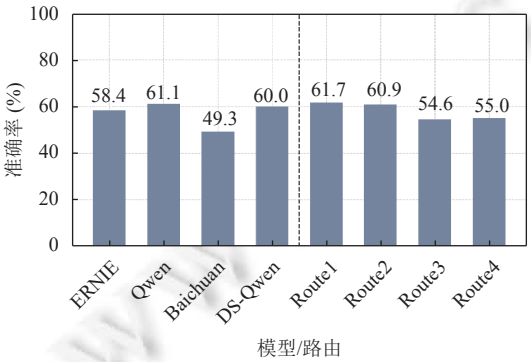

图 8 CaLM Lite 数据集准确率实验结果

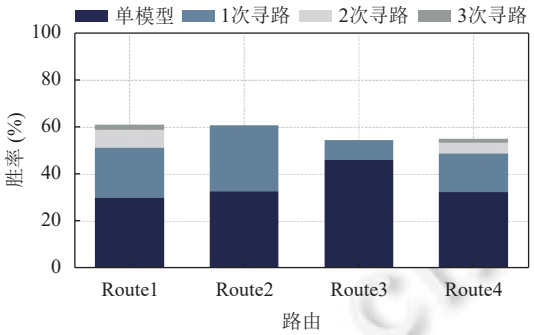

图 9 CaLM Lite 数据集路由情况

4.7 寻路次数分析

从 3 个数据集的实验结果可以看出, Route1 与 Route2 在大多数情况下均能在 3 次寻路以内完成路由. 然而, 当强制设置更少或更多的寻路次数时, 其对性能的具体影响尚不明确, 相较于本文提出的“自收敛机制结合寻路次数上限”策略, 仍需进一步实验验证与分析. 因此, 在不设置自收敛条件的前提下, 进行了寻路次数控制实验: 在无明确拒绝意见时, 随机选择合作模型继续优化, 从而与本文方法进行对比. 实验结果如表 4 和表 5 所示, 其中 Flames 无害率以及 Flames 无害分数为各维度结果求和取平均值.

在表 4 所示的 Route1 实验中, 由于初始模型 DS-Qwen 在 OMGEval 上性能起点较低, 随着寻路次数增加, 胜率显著提升. 然而, 在 Flames 指标上表现波动较大, 未表现出明显的提升趋势, 仅在一定范围内上下浮动. 在 CaLM Lite 上, 准确率随着寻路增加最初提升, 在中等寻路次数时达到峰值, 随后随着寻路次数进一步增加而出现性能退化. 表 5 展示了 Route2 的实验结果. 该结果表明 Route2 整体回复质量较高, 在 OMGEval 上的胜率还能够保持高位稳定. 然而, 其在 Flames 指标上同样表现出不稳定性, 且在 CaLM Lite 准确率上, 也呈现类似于 Route1 的“先升后降”的趋势. 综上所述, 在不引入自收敛机制的条件下, 仅通过控制寻路次数, 部分情况下仍能取得与“自收敛机制结合寻路次数上限”策略相当的性能表现. 例如, 表 4 中某些中间寻路次数配置在 CaLM Lite 上的表现与默认的自收敛寻路次数类似. 然而, 需注意的是, 自收敛策略下通常能在较少寻路次数内达成收敛 (例如, 3 次寻路以内路

由占比较高), 从而降低计算成本. 此外, 在表 4 与表 5 中的 Flames 与 CaLM Lite 结果还表明, 简单地增加寻路次数并不一定能够提升性能, 其提升幅度亦存在明显的上限.

表 4 寻路次数 α 的影响 (Route1)

寻路次数	OMGEval胜率 (%)	Flames 无害率 (%)	Flames 无害分数	CaLM Lite准确率 (%)
$\alpha=1$	59.7	27.5	61.50	60.7
$\alpha=2$	69.8	27.6	61.90	61.7
$\alpha=3$	75.3	27.9	61.93	61.9
$\alpha=4$	79.6	27.3	61.25	61.0
$\alpha=5$	79.6	25.3	60.16	60.1
$\alpha=6$	80.5	27.5	61.50	60.1

表 5 寻路次数 α 的影响 (Route2)

寻路次数	OMGEval胜率 (%)	Flames 无害率 (%)	Flames 无害分数	CaLM Lite准确率 (%)
$\alpha=1$	90.5	31.0	64.05	58.9
$\alpha=2$	90.0	32.6	65.06	59.1
$\alpha=3$	90.7	32.0	64.34	59.7
$\alpha=4$	91.0	32.1	64.37	59.9
$\alpha=5$	91.4	31.4	64.21	59.2
$\alpha=6$	91.8	33.2	65.17	58.6

5 总结与展望

为了细化模型串联路由质量评估策略、考虑模型之间的相互影响并提高路由选择效率, 本文在串联架构上设计了一种基于协作关系的模型动态路由方法. 该方法包含互评量化机制以及动态协作路由算法. 互评量化机制利用模型互评策略, 从各个模型的互评结果中挖掘协作关系信息, 为每个模型设定理想的协作对象. 动态协作路由算法则进一步利用协作关系信息, 以“一致同意规则”为基础, 从理想协作对象中挑选对当前输出结果的反对者作为下一跳节点. 大量实验结果表明本文所提路由算法的有效性. 互评分数与回复质量结果的相关性证实了分数与模型参与意愿、分数与模型协作关系的关联; 与对照组路由的比较则验证了该算法在回复质量上的显著优势. 根据实验结果可知, 大多数路由在两条之内即可结束, 且相较于随机设定的路由来说, 更能保证输出内容的质量.

通过显式建模模型间协作关系, 本文为模型互联引入了结构性协作的认知, 有助于推动从“系统整体协同”角度重新审视大模型集成方式; 通过引入动态路由机制, 本文实现了对多模型路径的按需调度与灵活组合, 为后续模型互联在复杂任务与环境下的部署提供了可行技术路线. 在未来工作中, 可以进一步探索多模型协作关系的动态演化机制, 支持更大规模、更复杂网络结构下的协作调度; 同时考虑引入用户评价与系统表现评估的反馈优化机制, 以提升路由策略的长期自适应能力.

References

[1] Brown T, Mann B, Ryder N, *et al*. Language models are few-shot learners. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1877–1901.

[2] Chen XK, Wu ZY, Liu XC, Pan ZZ, Liu W, Xie ZD, Yu XK, Ruan C. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. arXiv:2501.17811, 2025.

[3] Hadi MU, Tashi QA, Qureshi R, Shah A, Muneer A, Irfan M, Zafar A, Shaikh MB, Akhtar N, Hassan SZ, Shoman M, Wu J, Mirjalili S, Shah M. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. TechRxiv, 2024. [doi: 10.36227/techrxiv.23589741.v8]

[4] Gemini Team Google. Gemini: A family of highly capable multimodal models. arXiv:2312.11805, 2025.

[5] Wu XK, Chen M, Li WY, Wang R, Lu LM, Liu J, Hwang K, Hao YX, Pan YR, Meng QG, Huang KB, Hu L, Guizani M, Chao NP, Fortino G, Lin F, Tian YL, Niyato D, Wang FY. LLM fine-tuning: Concepts, opportunities, and challenges. Big Data and Cognitive

Computing, 2025, 9(4): 87. [doi: 10.3390/bdcc9040087]

[6] Chang YP, Wang X, Wang JD, Wu Y, Yang LY, Zhu KJ, Chen H, Yi XY, Wang CX, Wang YD, Ye W, Zhang Y, Chang Y, Yu PS, Yang Q, Xie X. A survey on evaluation of large language models. ACM Trans. on Intelligent Systems and Technology, 2024, 15(3): 39. [doi: 10.1145/3641289]

[7] Yu Y, Zhuang YC, Zhang JY, Meng Y, Ratner A, Krishna R, Shen JM, Zhang C. Large language model as attributed training data generator: A tale of diversity and bias. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc. , 2023. 55734–55784.

[8] Hajikhani A, Cole C. A critical review of large language models: Sensitivity, bias, and the path toward specialized AI. Quantitative Science Studies, 2024, 5(3): 736–756. [doi: 10.1162/qss_a_00310]

[9] Wang YB, Wang Y, Yu Y, Xu C, Yu H, Zhu ZL. Insights and analysis of open-source license violation risks in LLMs generated code. Ruan Jian Xue Bao/Journal of Software, 2025, 36(6): 2535–2557 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7324.htm [doi: 10.13328/j.cnki.jos.007324]

[10] Lin XY, Wang WJ, Li YQ, Yang S, Feng FL, Wei YW, Chua TS. Data-efficient fine-tuning for LLM-based recommendation. In: Proc. of the 47th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Washington: ACM, 2024. 365–374. [doi: 10.1145/3626772.3657807]

[11] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948, 2025.

[12] Yao HW, Lou J, Qin Z, Ren K. PromptCARE: Prompt copyright protection by watermark injection and verification. In: Proc. of the 2024 IEEE Symp. on Security and Privacy (SP). San Francisco: IEEE, 2024. 845–861. [doi: 10.1109/SP54263.2024.00209]

[13] Singhal K, Tu T, Gottweis J, *et al*. Toward expert-level medical question answering with large language models. Nature Medicine, 2025, 31(3): 943–950. [doi: 10.1038/s41591-024-03423-7]

[14] Gao YF, Yu DQ, Wang SQ, Wang HF. Large language model powered site selection recommender system. Journal of Computer Research and Development, 2024, 61(7): 1681–1696 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.202330629]

[15] Hazell J. Spear phishing with large language models. arXiv:2305.06972, 2023.

[16] Yuan Z, Yuan HY, Li CP, Dong GT, Lu KM, Tan CQ, Zhou C, Zhou JR. Scaling relationship on learning mathematical reasoning with large language models. arXiv:2308.01825, 2023.

[17] He XW, Lin ZH, Gong YY, Jin AL, Zhang H, Chen L, Jiao J, Yiu SM, Duan N, Chen WZ. AnnoLLM: Making large language models to be better crowdsourced annotators. In: Proc. of the 2024 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 6 (Industry Track). Mexico City: ACL, 2024. 165–190. [doi: 10.18653/v1/2024.naacl-industry.15]

[18] Lan YQ, Rao Y, Li GC, Sun L, Xia BC, Xin TT. A survey of text generation and evaluation based on intrinsic quality constraints. Acta Electronica Sinica, 2024, 52(2): 633–659 (in Chinese with English abstract). [doi: 10.12263/DZXB.20230826]

[19] Wu YQ, Zhou SY, Liu YF, Lu WM, Liu XZ, Zhang YT, Sun CL, Wu F, Kuang K. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 12060–12075. [doi: 10.18653/v1/2023.emnlp-main.740]

[20] Barbosa R, Santos R, Novais P. Collaborative problem-solving with LLM: A multi-agent system approach to solve complex tasks using Autogen. In: Proc. of the 2025 Int'l Conf. on Highlights in Practical Applications of Agents, Multi-agent Systems, and Digital Twins: The PAAMS Collection. Salamanca: Springer, 2025. 203–214. [doi: 10.1007/978-3-031-73058-0_17]

[21] Lan XC, Gao C, Jin DP, Li Y. Stance detection with collaborative role-infused LLM-based agents. In: Proc. of the 18th Int'l AAAI Conf. on Web and Social Media. Buffalo: AAAI, 2024. 891–903. [doi: 10.1609/icwsm.v18i1.31360]

[22] Weng YX, Wu GQ, Zheng TY, Yang YB, Luo J. Large model for small data: Foundation model for cross-modal RF human activity recognition. In: Proc. of the 22nd ACM Conf. on Embedded Networked Sensor Systems. Hangzhou: ACM, 2024. 436–449. [doi: 10.1145/3666025.3699349]

[23] Ge B, He CH, Zhang C, Xu H, Hu SZ. Unified transformation framework of open tag feature for Chinese multi-modal data. In: Proc. of the 9th Int'l Conf. on Big Data and Information Analytics (BigDIA). Haikou: IEEE, 2023. 692–698. [doi: 10.1109/BigDIA60676.2023.10429510]

[24] Zhang Y, Feng FL, Zhang JZ, Bao KQ, Wang QF, He XN. CoLLM: Integrating collaborative embeddings into large language models for recommendation. IEEE Trans. on Knowledge and Data Engineering, 2025, 37(5): 2329–2340. [doi: 10.1109/TKDE.2025.3540912]

[25] Tang YW, Zhang R, Liu JM, Guo Z, Zhao B, Wang ZG, Gao P, Li HS, Wang D, Li XL. Any2Point: Empowering any-modality large models for efficient 3D understanding. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2025. 456–473. [doi: 10.1007/978-3-031-72764-1_26]

[26] Jiang JC, Li Y, Nie J, Li JB, Wen BQ, Gadekallu TR. Integrating large language models with cross-modal data fusion for advanced intelligent transportation systems in sustainable cities development. Applied Soft Computing, 2025, 177: 113278. [doi: 10.1016/j.asoc.2025.113278]

[27] Chen C, Yin YC, Shang LF, Jiang X, Qin YJ, Wang FY, Wang Z, Chen X, Liu ZY, Liu Q. bert2BERT: Towards reusable pretrained language models. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Dublin: ACL, 2022. 2134–2148. [doi: 10.18653/v1/2022.acl-long.151]

[28] Xiao CJ, Zhang ZY, Song CY, *et al*. Configurable foundation models: Building LLMs from a modular perspective. arXiv:2409.02877, 2024.

[29] Wang S, Wang CH, Gao J, Qi Z, Zheng HY, Liao XX. Feature alignment-based knowledge distillation for efficient compression of large language models. arXiv:2412.19449, 2024.

[30] Ainsworth SK, Hayase J, Srinivasa S. Git re-basin: Merging models modulo permutation symmetries. arXiv:2209.04836, 2023.

[31] Liu TL, Guo SM, Bianco L, Calandriello D, Berthet Q, Llinares F, Hoffmann J, Dixon L, Valko M, Blondel M. Decoding-time realignment of language models. In: Proc. of the 41st Int'l Conf. on Machine Learning. Vienna: JMLR.org, 2024. 31015–31031.

[32] Lu P, Peng BL, Cheng H, Galley M, Chang KW, Wu YN, Zhu SC, Gao JF. Chameleon: Plug-and-play compositional reasoning with large language models. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 43447–43478.

[33] Aggarwal P, Madaan A, Anand A, Potharaju SP, Mishra S, Zhou P, Gupta A, Rajagopal D, Kappaganthu K, Yang YM, Upadhyay S, Faruqui M, Mausam. AutoMix: Automatically mixing language models. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2025. 131000–131034.

[34] Zhang MJ, Shen XM, Cao JN, Cui ZY, Jiang S. EdgeShard: Efficient LLM inference via collaborative edge computing. IEEE Internet of Things Journal, 2025, 12(10): 13119–13131. [doi: 10.1109/JIOT.2024.3524255]

[35] Jiang DF, Ren X, Lin BY. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Toronto: ACL, 2023. 14165–14178. [doi: 10.18653/v1/2023.acl-long.792]

[36] Brown B, Juravsky J, Ehrlich R, Clark R, Le QV, Ré C, Mirhoseini A. Large language monkeys: Scaling inference compute with repeated sampling. arXiv:2407.21787, 2024.

[37] Zhang YS, Sun RX, Chen YF, Pfister T, Zhang R, Arık SÖ. Chain of agents: Large language models collaborating on long-context tasks. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2025. 132208–132237.

[38] Zhao J, Zu C, Hao X, Lu Y, He W, Ding YW, Gui T, Zhang Q, Huang XJ. LONGAGENT: Achieving question answering for 128k-Token-Long documents through multi-agent collaboration. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: ACL, 2024. 16310–16324. [doi: 10.18653/v1/2024.emnlp-main.912]

[39] Sun XF, Li XY, Zhang SY, Wang SH, Wu F, Li JW, Zhang TW, Wang GY. Sentiment analysis through LLM negotiations. arXiv:2311.01876, 2023.

[40] Bo XH, Zhang ZY, Dai QY, Feng XY, Wang L, Li R, Chen X, Wen JR. Reflective multi-agent collaboration based on large language models. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2025. 138595–138631.

[41] Wang LQ, Zhou YT, Zhuang HY, Li QS, Cui D, Zhao YT, Wang L. Unity is strength: Collaborative LLM-based agents for code reviewer recommendation. In: Proc. of the 39th IEEE/ACM Int'l Conf. on Automated Software Engineering. Sacramento: ACM, 2024. 2235–2239. [doi: 10.1145/3691620.3695291]

[42] Ishibashi Y, Nishimura Y. Self-organized agents: A LLM multi-agent framework toward ultra large-scale code generation and optimization. arXiv:2404.02183, 2024.

[43] Wang XY, Chen YY, Yuan LF, Zhang YZ, Li YZ, Peng H, Ji H. Executable code actions elicit better LLM agents. In: Proc. of the 41st Int'l Conf. on Machine Learning. Vienna: JMLR.org, 2024. 50208–50232.

[44] Hari SN, Thomson M. Tryage: Real-time, intelligent routing of user prompts to large language models. arXiv:2308.11601, 2023.

[45] Kim S, Suk J, Longpre S, Lin BY, Shin J, Welleck S, Neubig G, Lee M, Lee K, Seo M. Prometheus 2: An open source language model specialized in evaluating other language models. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: ACL, 2024. 4334–4353. [doi: 10.18653/v1/2024.emnlp-main.248]

[46] Ye S, Kim D, Kim S, Hwang H, Kim S, Jo Y, Thorne J, Kim J, Seo M. FLASK: Fine-grained language model evaluation based on alignment skill sets. arXiv:2307.10928, 2024.

[47] Chen SH, Jiang WS, Lin BJ, Kwok JT, Zhang Y. RouterDC: Query-based router by dual contrastive learning for assembling large

language models. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2025. 66305–66328.

[48] Xu L, Lu XJ, Yuan CY, Zhang XW, Xu HL, Yuan H, Wei GA, Pan X, Tian X, Qin LB, Hai H. FewCLUE: A Chinese few-shot learning evaluation benchmark. arXiv:2107.07498, 2021.

[49] Liu Y, Xu M, Wang S, Yang L, Wang HY, Liu ZH, Kong CL, Chen Y, Liu Y, Sun MS, Yang EH. OMGEval: An open multilingual generative evaluation benchmark for large language models. arXiv:2402.13524, 2024.

[50] Huang KX, Liu XY, Guo QY, Sun TX, Sun JW, Wang YR, Zhou ZY, Wang YX, Teng Y, Qiu XP, Wang YC, Lin DH. Flames: Benchmarking value alignment of LLMs in Chinese. In: Proc. of the 2024 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers). Mexico City: ACL, 2024. 4551–4591. [doi: 10.18653/v1/2024.naacl-long.256]

[51] Chen SR, Peng B, Chen MQ, Wang RQ, Xu MY, Zeng XY, Zhao R, Zhao SJ, Qiao Y, Lu CC. Causal evaluation of language models. arXiv:2405.00622, 2024.

附中文参考文献

[9] 王毅博, 王莹, 余跃, 许畅, 于海, 朱志良. 大模型生成代码的开源许可证违规风险洞察与分析. 软件学报, 2025, 36(6): 2535–2557. http://www.jos.org.cn/1000-9825/7324.htm [doi: 10.13328/j.cnki.jos.007324]

[14] 高云帆, 郁董卿, 王思琪, 王昊奋. 大语言模型驱动的选址推荐系统. 计算机研究与发展, 2024, 61(7): 1681–1696. [doi: 10.7544/issn1000-1239.202330629]

[18] 兰玉乾, 饶元, 李冠呈, 孙菱, 夏昺灿, 辛婷婷. 基于内在质量约束的文本生成和评价综述. 电子学报, 2024, 52(2): 633–659. [doi: 10.12263/DZXB.20230826]

作者简介

吴俊儒, 博士生, 主要研究领域为人工智能, 大语言模型, 差分隐私.

李哲涛, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为人工智能, 物联网, 网络空间安全.

王建辉, 博士生, 主要研究领域为边缘计算, 人工智能.

刘忠仁, 博士生, CCF 学生会员, 主要研究领域为人工智能, 大语言模型推理系统.

庞永浩, 硕士生, 主要研究领域为大语言模型.

黄纪俊, 硕士生, 主要研究领域为大语言模型, 网络空间安全.