软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007702] [CSTR: 32375.14.jos.007702]

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

模型互联网中基于置信度的 Token 级并联协作推理*

王建辉, 李哲涛, 刘忠仁, 肖 勇

(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: 在模型互联网中, 单个基模型受限于自身推理能力, 其推理精度往往存在瓶颈. 虽然基模型可以通过与协作模型的 Token 级并联协作提升精度, 但现有方法多采用全程 Token 级并联协作策略, 导致较高的 Token 开销和推理时延. 针对上述挑战, 提出一种基于置信度的协作推理方法 CobCo. 该方法以基模型置信度为核心调控信号, 动态调节其推理行为. 当基模型置信度较低时, 引入协作推理以降低推理错误风险, 而在置信度较高时采用独立推理以减少冗余协作开销. 为支撑上述推理调节机制, 设计一种面向 Token 级并联协作的置信度评估算法, 通过将协作结果作为外部参照刻画基模型在逐步推理过程中的决策优劣, 并基于 Beta 分布对其置信度进行动态量化. 实验结果表明, 在引入 CobCo 后, 基模型的推理准确率相较于其独立推理提升了 2.4%–27.6%. 与现有全程 Token 级并联协作方法相比, CobCo 在推理准确率损失较小的前提下显著降低了 Token 开销与推理时延. 特别地, 当基模型与协作模型性能相近时, CobCo 能在保持推理准确率不下降的情况下, 将 Token 开销最高降低约 20.8%.

关键词: 大语言模型; 模型互联网; Token 级并联协作; 模型置信度; Beta 分布

中图法分类号: TP18

中文引用格式: 王建辉, 李哲涛, 刘忠仁, 肖勇. 模型互联网中基于置信度的Token级并联协作推理. 软件学报. http://www.jos.org.cn/1000-9825/7702.htm

英文引用格式: Wang JH, Li ZT, Liu ZR, Xiao Y. Confidence-aware Token-level Parallel Collaborative Inference in AI-model Network. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7702.htm

Confidence-aware Token-level Parallel Collaborative Inference in AI-model Network

WANG Jian-Hui, LI Zhe-Tao, LIU Zhong-Ren, XIAO Yong

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

Abstract: The inference accuracy of a single base model is often constrained by its reasoning capacity within an AI-model network. Although Token-level parallel collaboration with collaborative models can effectively improve inference accuracy, existing approaches generally rely on full-process Token-level parallel collaboration, resulting in substantial Token overhead and inference latency. To address these challenges, this study proposes a confidence-aware collaborative inference method, named CobCo. The proposed method leverages the base model's confidence as a core control signal to regulate its inference behavior dynamically. When the base model has low confidence, collaborative inference is activated to reduce the risk of errors. Conversely, when confidence is sufficiently high, the base model switches to independent inference to reduce redundant collaborative interactions. To support this adaptive inference mechanism, a confidence evaluation algorithm tailored for Token-level parallel collaboration is designed. By treating collaborative outcomes as external references, the proposed method characterizes the decision quality of the base model during step-by-step inference and dynamically quantifies its confidence using a Beta distribution. Experimental results show that integrating CobCo improves the base model's inference accuracy by 2.4% to 27.6% compared with independent inference. Compared with existing full-process Token-level parallel collaboration methods, CobCo significantly reduces Token overhead and inference latency while incurring only a small loss in inference accuracy. Notably, when the base model and collaborative models exhibit comparable performance, CobCo achieves up to a 20.8% reduction in

* 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); 广东省基础与应用基础研究重大项目 (2026B0303000007)
收稿时间: 2026-01-26; 修改时间: 2026-03-17, 2026-04-28; 采用时间: 2026-05-18; jos 在线出版时间: 2026-08-26

Token overhead without any degradation in inference accuracy.

Key words: large language model (LLM); AI-model network; Token-level parallel collaboration; model confidence; Beta distribution

随着大模型技术的发展, 其推理能力不断增强, 然而单一模型在复杂任务中仍存在性能局限, 多模型协作逐渐成为提升推理能力的重要手段[1−3]. 在此背景下, 一种以多模型协作为核心的模型互联网 (AI-model network) 逐渐受到关注[1]. 该网络由海量部署在边端的模型构成, 并依托统一的模型互联平台实现模型调用、协作推理及任务调度, 其核心目标在于构建开放、动态的模型互联生态. 在该网络中, 任务的发起者称为基模型, 其余参与协作的模型为协作模型, 二者通过结构化的协作流程相互配合以增强复杂问题的求解能力[4−6].

从协作粒度来看, 现有多模型协作方法主要分为全回复级协作[7]和 Token 级并联协作两类[8]. 全回复级协作通过整合多个模型的完整回复提升输出质量, 但在复杂任务中通常依赖多轮整合, 带来显著的计算开销与推理时延. Token 级并联协作则以更细粒度的方式逐步聚合候选 Token, 在单轮推理内完成决策, 兼顾效率与性能, 成为当前多模型协作推理的研究热点. 针对 Token 级并联协作的实现方式, 现有方法按照聚合策略分为概率层聚合和逻辑值层聚合两类. 前者指在聚合时采用归一化后的概率值向量, 如 GaC[9]、UniTe[10]. 而后者指采用未归一化的逻辑值向量进行聚合以保留不同模型的输出尺度区别, 如 DuetNet[8].

如图 1 所示, 基模型独立推理在推理效率方面具有优势, 但准确率受模型能力所限; 相比之下, Token 级并联协作推理则通过聚合多模型共识降低推理错误风险, 在准确率上表现更优. 然而, 当前 Token 级并联协作方法普遍局限于以推理准确率为单一优化目标, 同时依赖于低效的全程 Token 级并联协作. 该策略存在两点主要局限: (1) 全程 Token 级并联协作未能有效利用基模型的独立推理能力. 无论任务难易, 强制全程 Token 级并联协作将导致不必要的 Token 开销激增. (2) 该策略会放大 Token 生成的“短板效应”. 系统需要集成所有模型的输出分布以确定下一个 Token, 因此单个模型的时延 (如因算力或通信问题) 将制约整体生成速度. 在全程 Token 级并联协作中, 每一个 Token 均面临此类风险, 显著增加任务总耗时.

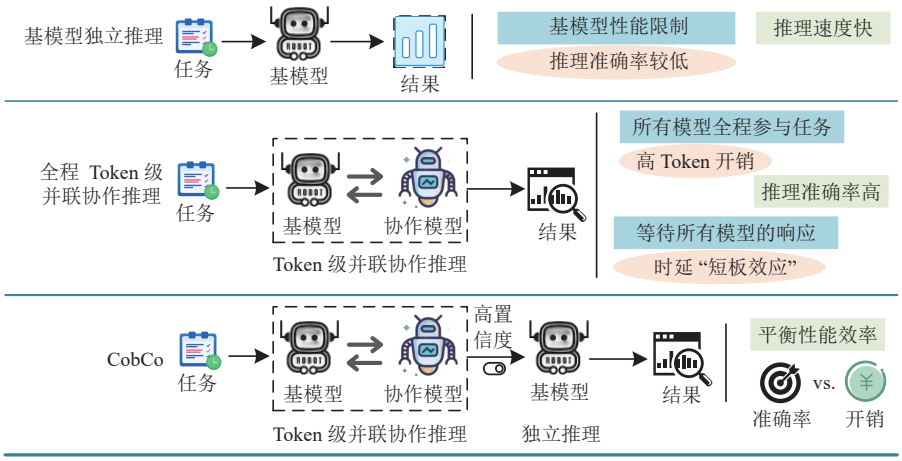


图 1 基模型独立推理、全程 Token 级并联协作推理和 CobCo 方法对比

上述协作开销累积和时延受限的问题, 在一定程度上源于现有方法缺乏对协作过程进行自适应调节的能力, 即难以根据当前推理状态判断协作是否仍然必要. 直观而言, 当模型在当前上下文下已经具有较高把握时, 继续进行 Token 级并联协作可能带来额外开销而收益有限, 而此时转为独立推理往往能够在保持较高准确率的同时降低协作开销. 为刻画这种“把握程度”, 本文引入置信度, 用于反映模型对其后续推理正确性的信心. 基于此, 本文提出一种基于置信度的 Token 级并联协作推理方法 CobCo. 该方法将推理过程划分为 Token 级并联协作推理与独立推理两个阶段, 并以置信度为依据进行动态调节. 推理初期采用基模型与协作模型的 Token 级并联协作, 当置信度较高时切换至基模型独立推理, 否则维持协作过程. 相较于全程 Token 级并联协作方法, CobCo 通过适时转入独立

推理有效降低了总体 Token 开销; 同时, 逐 Token 协作次数的减少缓解了由模型响应差异引发的时延短板效应, 从而提升整体推理效率.

然而, CobCo 方法中动态调节机制的有效性依赖于对基模型置信度的准确刻画. 若置信度评估偏高, 基模型可能过早进入独立推理阶段, 从而影响结果准确性; 反之, 若置信度评估偏低, 则会导致基模型过晚转入独立推理, 从而限制效率提升. 因此, CobCo 的一个核心设计难点, 即在于对基模型置信度的准确评估. 受不确定性认同理论启发[11], 本文提出一种基于 Beta 分布的置信度评估方法, 从而为基模型动态转入独立推理提供合理的决策依据. 该理论揭示, 个体信心通过在群体中与同伴的社会比较而动态形成: 表现优于同伴则信心增强, 反之则减弱. 本文将协作推理视为一类群体社会行为, 在每一步 Token 生成中, 若基模型表现优于协作模型, 则视为一次正反馈, 增强其置信证据; 反之则为负反馈, 削弱置信证据. 该二项式证据更新机制在数学形式上与 Beta 分布作为二项分布共轭先验的性质同构, 使得置信度估计可自然嵌入 Beta 分布的统计框架. 本文的主要贡献如下.

1) 提出一种基于置信度的 Token 级并联协作推理方法 CobCo. 该方法以基模型为决策主体, 通过置信度驱动协作推理与独立推理之间的动态切换, 在保留 Token 级并联协作提升复杂任务推理准确率优势的同时, 避免不必要的全程 Token 级并联协作, 从而在准确率与推理效率之间实现有效平衡.

2) 设计一种面向协作推理的置信度建模算法, 为 CobCo 实现自适应的精准推理阶段切换提供关键支撑. 该算法将 Token 级并联协作过程中生成的共识 Token 在基模型与协作模型的相对优劣作为外部反馈信号, 并基于 Beta 分布对基模型推理置信度进行概率建模与动态更新.

3) 实验结果表明, CobCo 在显著提升基模型推理准确率的同时, 展现出优异的效率优势. 在基模型与协作模型性能相近的单协作场景中, CobCo 可在不降低准确率的情况下, 将 Token 开销最高降低约 20.8%. 在多协作模型场景中, 与全程 Token 级并联协作方法相比, CobCo 的平均推理准确率最大降幅约 7.2%, 但可将 Token 开销最高降低 33.4%, 单 Token 生成时延最高减少 34%.

本文第 1 节描述相关工作. 第 2 节详细介绍本文提出的 CobCo 方法, 包括基于置信度的 Token 级并联协作推理流程和置信度评估算法. 第 3 节通过实验验证所提方法的有效性. 第 4 节总结全文.

1 相关工作

多模型协作是一种通过整合多个模型的推理能力以协同完成同一任务的计算范式, 其核心目标在于通过模型间的优势互补来提升整体性能与问题解决能力[6−8]. 根据协作过程中信息聚合的粒度差异, 现有多模型协作方法分为全回复级协作和 Token 级并联协作.

全回复级协作通过对多个模型生成的完整回复进行集成与融合, 以获得统一且性能更优的最终输出结果. 混合智能体 (mixture-of-agents, MoA) 是一类采用全回复级协作机制的典型框架[12]. MoA 框架采用分层设计, 每层基于上一层的结果构建, 逐步提升回答质量. Yu 等人[13]提出了一种多智能体协作框架 MACT, 通过规划、执行、判断和答案这 4 类智能体的分工协作完成任务. Tao 等人[14]提出了讨论链, 利用大模型协作回答复杂问题. 除通用领域外, 多模型协作亦广泛应用于多个细分领域, 如软件开发[15]、运筹学任务[16]等. 不同于侧重机制设计的研究, Qian 等人[17]研究了协作智能体数量对性能的影响, 并发现性能随智能体数量增加而提升.

Token 级并联协作是一类推理时聚合方法, 其通过对不同模型产生的候选 Token 概率分布进行聚合, 实现逐步的 Token 生成. Yu 等人[9]提出了一种基于全量聚合策略的 Token 级并联协作方法 GaC. 该方法通过聚合多个模型输出的全部候选 Token 实现 Token 级并联协作推理. 为降低计算开销和减少聚合过程的噪声, Yao 等人[10]提出了 UniTe, 它仅关注每个模型输出分布中的 Top-K 部分. 为进一步优化结果, 本文作者此前工作[8]设计了一种联合 Top-K 和 Top-P 截断的 Token 级并联协作方法 DuetNet, 并且其采用聚合逻辑值向量以减少归一化处理导致的置信度损失. 不同于上述等权重聚合方法, Mavromatis 等人[18]提出了一种加权平均方法, 其中权重由每个模型在特定生成步骤的困惑度所决定. Jin 等人[19]和 Li 等人[20]采用同质模型进行 Token 级并联协作从而避免了对齐操作. Jin 等人[21]提出了一种基于模型置信度的路由方法 CE-CoLLM, 该方法根据置信度在基模型与协作模型之间进行

切换, 以提升基模型的推理表现. 然而, 由于其属于路由范式, 协作模型的参与依赖于基模型的间歇式触发机制, 在一定程度上可能引入额外的调度与响应开销.

上述研究表明, 全回复级协作方法往往需要多轮完整响应交互, 计算成本较高. 相比之下, Token 级并联协作仅需单轮响应即可完成, 推理效率更高. 然而, 当前 Token 级并联协作研究主要采用全程 Token 级并联协作模式, 即所有参与模型在整个推理过程中持续保持协同工作状态. 这种协作方式虽然能确保模型间的深度交互, 但仍存在以下局限性: 一是计算资源消耗大, 所有模型需全程保持活跃状态; 二是缺乏动态适应性, 无法根据任务需求灵活调整协作强度; 三是全程 Token 级并联协作使得“短板效应”加剧, 由于 Token 级并联协作的生成速率受限于最慢的参与模型, 全程 Token 级并联协作模式会持续放大这一性能瓶颈, 导致整体推理效率降低. 因此, 针对现有方法的不足, 本文设计了一种基于置信度的协作推理方法.

2 基于置信度的 Token 级并联协作推理方法 CobCo

本节详细描述了基于置信度的协作推理方法 CobCo. 首先概述 CobCo 方法, 然后介绍基于置信度的 Token 级并联协作推理算法, 最后描述基模型置信度评估算法. 表 1 列出了本文中使用的主要符号及其含义.

表 1 主要符号及其含义

符号	含义
M	参与协作的模型数量
m	第 m 个模型, 其中 $m=1$ 表示基模型, $m>1$ 表示协作模型
logits_m	给定输入文本 I 的条件下模型 m 输出的逻辑值向量
$\text{logits}_m[u]$	在 logits_m 中, Token u 对应的逻辑值
$G_m(\cdot)$, V_m, $\lvert V_m\rvert$	分别表示模型 m 对应的生成函数、词汇表和词汇表大小
cand_m	对模型 m 的输出进行联合截断后的候选Token集合
L_u	候选Token $u \in U$ 的累计逻辑概率值
U, L	分别表示所有模型输出的候选Token集合的并集及对应的累计逻辑值集合
s_u^m	Token u 在模型 m 的候选集合中逻辑值大小
u_{next}	在聚合结果的基础上采样得到下一个Token
W	观察窗口的大小
K, P	分别表示联合截断策略中Top-K取值和Top-P的取值
φ	置信度阈值
s, f	分别表示正向反馈计数与负向反馈计数
ρ	惩罚系数
T	基模型置信度估计值
$X=\{x_1,x_2,\ldots,x_\tau\}$	观测到的 τ 次独立的置信度证据集合

2.1 CobCo 方法概述

本节首先简述传统的 Token 级并联协作推理, 然后概述 CobCo 方法.

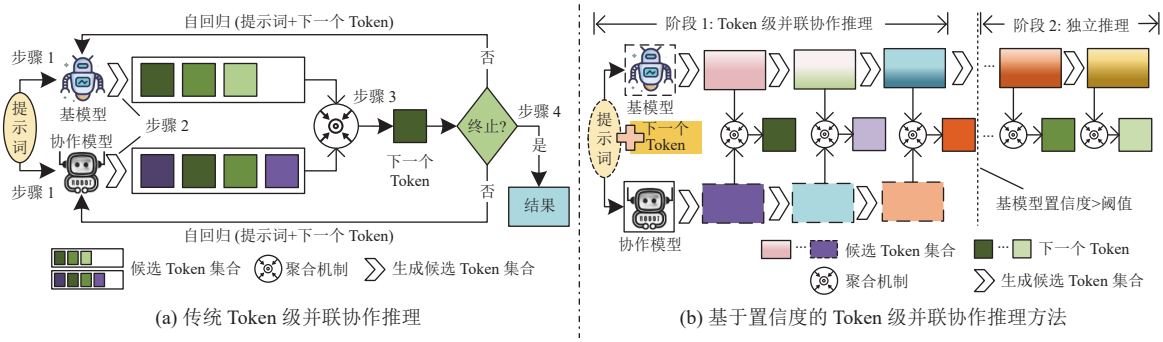
图 2(a) 展示了传统的 Token 级并联协作推理的过程, 该过程包括以下 4 个步骤.

步骤 1: 输入处理. 每个参与协作的模型首先利用自身的分词器将输入文本拆分为离散的 Token 序列, 并根据其词汇表将这些 Token 映射为相应的索引 (即 Token ID), 从而得到模型可处理的数值化输入序列.

步骤 2: 候选生成. 每个模型基于输入序列在设定的生成规则下输出一组候选 Token 集合, 例如选取概率最高的 K 个候选 Token.

步骤 3: 候选聚合. 通过某种聚合机制聚合多个协作模型生成的候选 Token 集合得到综合候选集, 紧接着基于采样规则从综合候选集选择出一个 Token 作为每个模型共用的下一个 Token.

步骤 4: 序列更新与终止判定. 将该下一个 Token 附加到每个协作模型的输入序列中; 重复步骤 2–4, 直到满足终止条件, 例如, 输出终止符或达到最大生成 Token 数量.



(a) 传统 Token 级并联协作推理　　(b) 基于置信度的 Token 级并联协作推理方法

图 2　传统 Token 级并联协作推理方法与 CobCo 方法

定义 1. 基模型置信度. 基模型置信度指基模型量化其自身能够独立完成当前推理任务并输出正确结果的置信水平. 该置信水平越高, 表明基模型独立正确推理的可能性越高.

本文提出的 CobCo 方法如图 2(b) 所示. 在推理流程上, CobCo 将任务执行划分为两个阶段: Token 级并联协作推理阶段和独立推理阶段. 在 Token 级并联协作推理阶段, 基模型与协作模型通过 Token 级并联协作生成后续 Token, 形成协作推理序列; 若基模型的置信度超过预设阈值, 则切换至独立推理阶段, 基模型将在已有协作序列基础上继续自主推理. 反之, 若置信度未达阈值, 系统将维持 Token 级并联协作推理阶段, 基模型继续借助协作模型进行协同生成.

CobCo 的基本原理源于社会心理学理论中的不确定性认同理论[11]. 该理论指出个体在面对不确定性决策时, 会通过增强群体认同来减少不确定性, 从而缓解焦虑和增强归属感. 相应地, 在 Token 级并联协作推理阶段, 基模型通过与其他模型进行持续的 Token 级交互, 不断吸收集体推理中的共识信息, 从而逐步降低自身在任务推理中的不确定性. 这种多模型协同的工作机制, 在功能上类似于个体通过群体认同获取认知确认的过程. 进一步, 当基模型通过协作使其不确定性显著降低后, 便具备了独立决策的认知条件, 此时系统自然过渡到独立推理阶段. 这一阶段对应了个体在获得足够认知确信后脱离群体进行自主决策的心理过程.

2.2　基于置信度的 Token 级并联协作推理

基于置信度的 Token 级并联协作推理包含两个核心阶段: Token 级并联协作推理与独立推理. 假设共有 M 个模型参与 Token 级并联协作推理, 其中 $m \in \{1\}$ 表示基模型, $m \in \{2,3,\ldots,M\}$ 表示协作模型 m. 在 Token 级并联协作推理阶段, 每个 Token 的生成过程包括 3 个子步骤: 候选 Token 集合生成、候选 Token 集合聚合以及聚合结果采样, 具体如下.

1) 候选 Token 集合生成. 设 I 表示用户输入的提示词文本, 则模型 m 在给定输入 I 的条件下输出的逻辑值向量 logits_m 表示为:

$$\text{logits}_m = G_m(I) \tag{1}$$

其中, $G_m(\cdot)$ 表示模型 m 对应的生成函数, 其输出为未进行归一化的对数概率向量; logits_m 向量的大小为模型 m 的词汇表大小 $|V_m|$, 该向量中的每个元素的值表示对应候选 Token 的未归一化得分. Token $u \in V_m$ 的逻辑值 $\text{logits}_m[u]$ 反映的是对数概率, 该值越大则表示 Token u 被选中的概率越高.

随后, 对 logits_m 进行截断以过滤低置信度噪声. 在 CobCo 中, 采用联合 Top-K 截断和 Top-P 截断的策略以过滤噪声. 具体而言, 先通过 Top-K 保留概率最高的 K 个候选 Token, 再基于 Top-P 从中选取累积概率达到阈值 P 的最小候选集, 从而兼顾生成质量与稳定性. 那么, 联合截断后的候选 Token 集合 cand_m 可表示为:

$$\text{cand}_m = F(\text{logits}_m, K, P) \tag{2}$$

其中, $F(\cdot)$ 表示对 logits_m 进行联合截断.

2) 候选 Token 集合聚合. 将候选 Token 集合进行逐 Token 累加聚合, 得到 Token $u \in U$ 的累计逻辑概率值 L_u:

$$L_u = \sum_{m=1}^{M} s_u^m,\ u \in U \tag{3}$$

其中, U 表示所有模型输出的候选 Token 集合的并集; $L_u \in L$, L 为集合 U 对应的累计逻辑值集合; s_u^m 表示 Token u 在模型 m 的候选集合中逻辑值大小, 其可通过公式 (4) 得到.

$$s_u^m = \begin{cases} 0, & u \notin \text{cand}_m \\ \text{logits}_m[u], & u \in \text{cand}_m \end{cases} \tag{4}$$

3) 聚合结果采样. 在聚合结果的基础上采样得到下一个 Token u_{next}, 即共识 Token. 聚合后候选 Token 的逻辑值越高表示该 Token 的置信度越强, 能更好地反映多个模型的共识. 因此, 与现有研究[8,10]相似, 本文采用贪心策略进行采样, 即通过公式 (5) 采样得到 u_{next}.

$$u_{\text{next}} = \arg\max_{u \in U} L_u \tag{5}$$

进一步地, 当基模型置信度超过预设阈值时, 系统将切换至独立推理阶段. 在此阶段, 由于协作模型不再参与, 此时推理过程可视为 Token 级并联协作推理的一种特例, 即参与协作的协作模型数量为零. 因此, 前述 Token 级并联协作推理流程同样适用于独立推理阶段.

基于置信度的 Token 级并联协作推理流程的伪代码如算法 1 所示. 算法输入提示词文本 I、观察窗口大小 W、联合截断策略中的 K 值和 P 值、置信度阈值 φ 以及每个模型的生成函数 $G_m(\cdot)$.

算法 1. 基于置信度的 Token 级并联协作推理算法.

输入: I, W, K, P, φ, $G_m(\cdot)$, $m = 1,2,\ldots,M$;
输出: 回答 A.

1. //初始时, Token 级并联协作推理阶段, $C = \{1\}$ 等价于独立推理
2. 初始化空字符串 A; $C = \{1,2,\ldots,M\}$;
3. **do**
4. **for** 模型 m **in** C **do**
5. $\text{logits}_m \leftarrow G_m(I)$;
6. $\text{cand}_m = F(\text{logits}_m, K, P)$;
7. **end for**
8. 统计得到综合候选集 U 及对应的累计逻辑值集合 L;
9. 从 U 中选出累计逻辑值最大的 Token u_{next};
10. $I \leftarrow I + u_{\text{next}}$; $A \leftarrow A + u_{\text{next}}$; $W \leftarrow W - 1$;
11. 采集用于评估基模型置信度的置信度证据数据;
12. 如果 $W \leqslant 0$, 则评估基模型的置信度 T;
13. //若置信度大于阈值, 则进入独立推理阶段
14. 如果基模型的置信度 T 大于阈值 φ, 则 $C \leftarrow \{1\}$;
15. **while** $u_{\text{next}} \neq$ "end" **and** Token 数量未超出
16. **return** A

在任务初始时, 推理过程进入 Token 级并联协作阶段, 基模型与协作模型通过 Token 级并联协作逐步生成后续 Token, 并同步采集用于评估基模型置信度的证据数据. 为避免因证据样本不足而导致置信度估计偏差, 本文引入观察窗口机制, 以在累积足够协作反馈后对基模型置信水平进行稳定评估. 具体而言, 如算法 1 的第 12 行所示, 当协作生成的 Token 数小于窗口大小 W 时, 暂不进行置信度评估; 否则评估当前步骤基模型的置信度 T.

由于独立推理过程可视为 Token 级并联协作推理的一种特例, 因此算法 1 中参与协作的模型集合 $C = \{1\}$ 表

示基模型独立推理. 此外, 基于置信度的 Token 级并联协作推理算法同样遵循自回归生成机制, 逐步构建推理结果. 该过程以已生成的 Token 序列为条件, 迭代地预测下一个 Token, 直至输出结束字符或者达到最大 Token 数量限制. 具体而言, 正如上文所述, 在每一个生成步中, 算法首先构建候选 Token 集合 (第 4–7 行), 随后对来自多模型的候选结果进行聚合 (第 8 行), 并通过采样策略确定最终输出的 Token (第 9 行). 上述过程在自回归框架下不断重复, 逐步扩展生成序列, 最终得到协作推理结果 A.

2.3 基模型置信度评估

本节详细阐述一种面向 Token 级并联协作推理的基模型置信度评估算法.

定义 2. 候选 Token 的优先级. 令 $m_t=\{m_1^t,m_2^t,\ldots,m_N^t\}$ 表示模型 m 在第 t 个生成步骤中生成的按概率降序排列的候选 Token 集合. 候选 Token 的优先级越高, 在采样过程中越容易被选中. 本文定义 Token m_i^t 的优先级为其在有序集合 m_t 的下标 i, 且下标越小优先级越高.

定义 3. 基模型相对于协作模型的单步推理优劣. 令 γ_t 表示第 t 个生成步骤中通过 Token 级并联协作推理生成的下一个 Token. A_t 和 B_t 分别表示基模型 α 和协作模型 β 在第 t 个生成步骤中生成的候选 Token 集合. 若协作推理生成的共识 Token γ_t 在基模型 α 中的优先级不低于协作模型, 则表明基模型的预测分布更近于群体共识, 可以认为基模型 α 优于协作模型 β; 反之, 则协作模型 β 更优. 该比较逻辑可形式化表示为公式 (6):

$$\begin{cases}\alpha\text{优于}\beta, \text{ if } (\gamma_t\in A_t \text{ and } \gamma_t\notin B_t) \text{ or } A_t^{\lambda}\leqslant B_t^{\lambda}\\ \alpha\text{劣于}\beta, \text{ if } (\gamma_t\notin A_t \text{ and } \gamma_t\in B_t) \text{ or } A_t^{\lambda}> B_t^{\lambda}\end{cases} \tag{6}$$

其中, A_t^{λ} 和 B_t^{λ} 分别表示 γ_t 在 A_t 和 B_t 中的优先级. 由公式 (6) 可见, 若 γ_t 仅存在于候选集合 A_t 或 γ_t 在 A_t 的优先级不低于在 B_t 的优先级, 那么基模型 α 在第 t 个生成步骤中优于协作模型 β. 反之, 若 γ_t 仅存在于协作模型候选集合 B_t, 或者 γ_t 在 A_t 的优先级低于在 B_t 的优先级, 那么基模型 α 在第 t 个生成步骤中劣于协作模型 β.

根据定义 3, 在单步生成过程中, 若基模型对下一个 Token 的预测优于协作模型, 则表明当前步骤的共识机制更倾向于基模型的输出. 换言之, 基模型在当前 Token 的协作生成结果中表现出比协作模型更高的置信度. 为便于理解, 图 3 结合具体示例进一步阐释了定义 3 中基模型与协作模型在单步生成中的相对优劣关系. 如图 3 所示, 在当前生成步骤中, 基模型和协作模型输出的候选 Token 集合分别为{“我是一个”, “我是”, “我”, “你好”, “您”}和{“我”, “你”, “您”, “I”, “assistant”}. 经聚合和采样步骤后, 最终生成的 Token 为“我”. 该 Token 在基模型的候选 Token 集合中的优先级为 2; 而在协作模型中的优先级为 0. 由于协作模型优先级更高, 因此判定基模型在当前生成步骤中劣于协作模型.

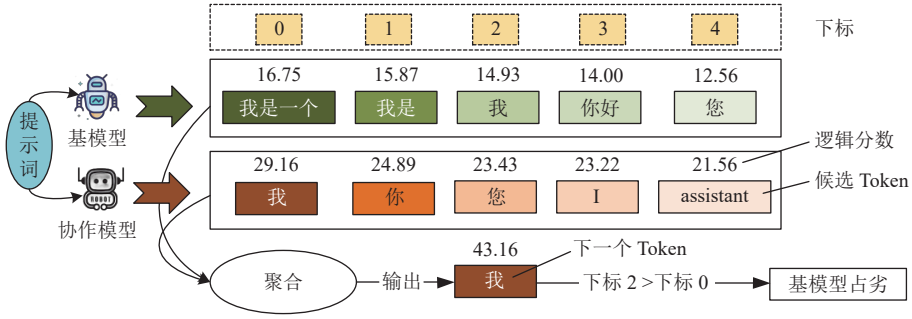


图 3 基模型与协作模型在一个生成步骤中的优劣判断

在 CobCo 中, 准确量化基模型的置信度至关重要, 它直接决定了模型在切换至独立推理阶段后能否维持正确的输出结果. 正如社会认同理论所揭示的个体在群体中的自信心是通过与同伴的持续社会比较而动态构建的. 当个体在某项任务中的表现优于同伴时, 其自信心会随之增强; 反之则会减弱. 这一过程本质上体现为基于历史交互反馈的信念更新机制.

将上述思想映射至 CobCo 方法, 在 Token 级并联协作的群体内, 基模型在每个 Token 的生成过程中都会与作

为同伴的协作模型进行局部性能比较. 具体而言, 若基模型在当前生成步骤优于协作模型, 则可视为一次正向反馈, 其置信度相应增加; 反之, 则视为负向反馈, 其置信度降低. 由此, 基模型的置信度可被视为一种随社会比较结果动态演化的自我效能表征. 进一步地, 鉴于单次比较结果具有二元性特征, 可将其建模为一次伯努利试验. 在此基础上, 本文采用 Beta 分布对基模型的置信度进行刻画, 该建模方式与其统计假设天然契合, 从而能够在贝叶斯框架下实现对置信度的动态更新与不确定性建模. 具体地, 将每次社会比较视为一次伯努利试验, 其中正向反馈记为成功, 负向反馈记为失败, 则经过多次交互后, 基模型的置信度可通过 Beta 分布的后验参数进行更新, 并由其期望形式进行度量. 由此, 社会认同理论中的“社会比较-信心演化”过程被形式化为“反馈累积-置信度更新”的概率建模机制, 使得所定义的置信度不仅具备直观解释性, 同时具备良好的可计算性与可更新性.

基于上述建模思想, 本文将基模型在每一 Token 生成步骤中与协作模型的优劣比较, 形式化为一次伯努利试验. 定义随机变量 $Y \in \{0,1\}$ 表示基模型在每一步 Token 生成中的优劣表现. 其中 $Y = 1$ 表示基模型在当前生成步骤优于协作模型, 即获得正向社会反馈, 视为一次成功; 相反 $Y = 0$ 表示基模型劣于协作模型, 视为一次失败. 进一步, 令 p 表示基模型潜在的真实置信度, 假设其先验分布服从参数为 $s > 0$ 和 $f > 0$ 的 Beta 分布, 即:

$$p \sim Beta(s, f) \tag{7}$$

其中, 形状参数 s 和 f 分别表示基模型的先验正向与负向反馈计数, 初始均设为 1.

令集合 $X = \{x_1, x_2, \ldots, x_\tau\}$ 表示观测到的 τ 次独立置信度证据, 其中 τ 表示当前生成的 Token 数量, x_t $(t = 1, 2, \ldots, \tau)$ 为第 t 个生成过程得到的置信度证据. 针对多协作模型场景, 本文对伯努利试验中的成功与失败计数进行了扩展定义. 如公式 (8) 所示, 当基模型不优于任意一个协作模型时, 视为失败, 并记录惩罚性失败计数 $\rho > 1$; 若基模型优于某个协作模型, 则视为一次成功, 且成功计数为基模型所优于的协作模型数量. 在多协作模型场景下, 扩展成功计数能更精确地量化基模型的相对表现, 而统一惩罚则在基模型表现全面劣势时提供强负反馈, 从而有效抑制基模型的过度自信.

$$x_t = \begin{cases} +\pi, & \text{基模型优于至少一个协作模型} \\ -\rho, & \text{否则} \end{cases} \tag{8}$$

其中, π 为第 t 个生成过程中劣于基模型的协作模型数量, $\rho \geqslant 1$ 为惩罚系数. 当 x_t 为正数时表示成功计数, 此时 s 累加 x_t; 反之表示失败计数, 此时 f 累加 $|x_t|$. 由此, 根据共轭性, p 的后验分布更新为:

$$P(p \mid X) = Beta\left(s + \sum_{x_t > 0} x_t, f + \sum_{x_t \leq 0} |x_t|\right) \tag{9}$$

由此, 基于 Beta 分布原理, 本文采用后验分布的期望值作为基模型置信度估计值 T, 即:

$$T = E[p \mid X] = \frac{s + \sum_{x_t > 0} x_t}{\left(s + \sum_{x_t > 0} x_t\right) + \left(f + \sum_{x_t \leqslant 0} |x_t|\right)} \tag{10}$$

由公式 (10) 可见, 估计值 T 随正向反馈的增加而增加, 随负向反馈的增加而下降, 这与社会认同理论的预期一致. 此外, 参数 s 和 f 的增量更新模拟了置信度社会互动中的累积过程. 最后, 整个更新过程仅为简单的计数操作, 计算开销极低.

为确定基模型在 Token 级并联协作与独立推理之间的切换时机, 本文采用基于置信度阈值的决策策略. 若当前置信度超过阈值 φ, 转入独立推理; 否则保持 Token 级并联协作推理. 基模型置信度评估算法如算法 2 所示.

算法 2. 基模型置信度评估算法.

输入: X, s, f;
输出: 基模型置信度 T.

1. 如果 $|X| < W$, 则返回初始置信度, 即 **return** 0;
2. **for** each $x_t \in X$ **do**

3. **if** $x_t > 0$ **then**
4. $s \leftarrow s + x_t$;
5. **else**
6. $f \leftarrow f + |x_t|$;
7. **end if**
8. **end for**
9. 计算基模型置信度 T, 即 $T \leftarrow s/(s+f)$;
10. **return** T

3 实验分析

3.1 实验设置

如表 2 所示, 本文选取了 5 个开源大语言模型进行实验以验证 CobCo 的有效性. 实验在配备 4 块 80 GB 显存的 A100 服务器上进行. 所有模型在解码阶段均采用统一的联合截断策略, 其中 Top-K 参数 K 设为 10, Top-P 参数 P 设为 0.75[8]. 本文采用表 3 所列的 5 种推理数据集对 CobCo 进行性能评估, 涵盖数学计算、选择题与判断题等多种题型. 本文从每类数据集中随机抽取 50 道题目, 构成统一的测试集. 为确保实验的公平性, 单模型独立推理与多模型 Token 级并联协作推理均采用完全相同的提示词模板.

表 2 实验采用的开源大模型

简称	模型名称	发布者
q1	Qwen1.5-7B-Chat[22]	阿里巴巴集团
q2	Qwen2.5-7B-Instruct[23]	
q2M	Qwen2.5-14B-Instruct[23]	
g4	glm-4-9B-chat[24]	智谱AI
L3	Meta-Llama-3.1-8B-Instruct[25]	Meta

表 3 实验采用的 5 种测试数据集及对应的提示词模板

数据集	数据类型	描述	提示词模板
TruthfulQA[26]	选择题	衡量语言模型在生成问题答案时是否真实的基准数据集	Can you answer the following question as accurately as possible? **{question}** You must put the answer (such as (A) or (B)) at the end of your response.
MMLU[27]	选择题	涵盖57个学科的多任务语言理解基准数据集	Can you answer the following question as accurately as possible? **{question}** You must put the answer (such as (A) or (B)) at the end of your response.
Math[28]	数学计算题	计算 $a+b*c+d-e*f$, 其中 $a-f$ 为0–30的随机整数	What is the result of $a+b*c+d-e*f$? You must reiterate your answer numerically at the end of the response.
GSM8K[29]	数学推理题	由8.5k高质量的小学数学文字题组成的数据集	Please answer the following question: **{question}** You must reiterate your answer numerically at the end of the response.
BoolQ[30]	判断题	用于回答是/否问题的英文数据集, 这些问题在无提示和无约束的环境中生成	Read the following background: **{background}** Based on the above background, please answer the True-false question: **{question}** If you think it is correct, answer (True), otherwise answer (False). You must reiterate your answer at the end of the question.

本文对比的方法如下.

单模型: 仅采用基模型独立推理的方法, 且其生成过程的步骤与 CobCo 一致.

DuetNet[8]: 一种采用联合截断策略的逻辑值向量维度的 Token 级并联协作方法. 该方法采用全程 Token 级并

联协作策略, 参照原论文设置, 本文将联合截断策略的 Top-K 和 Top-P 值分别设定为 10 和 0.75.

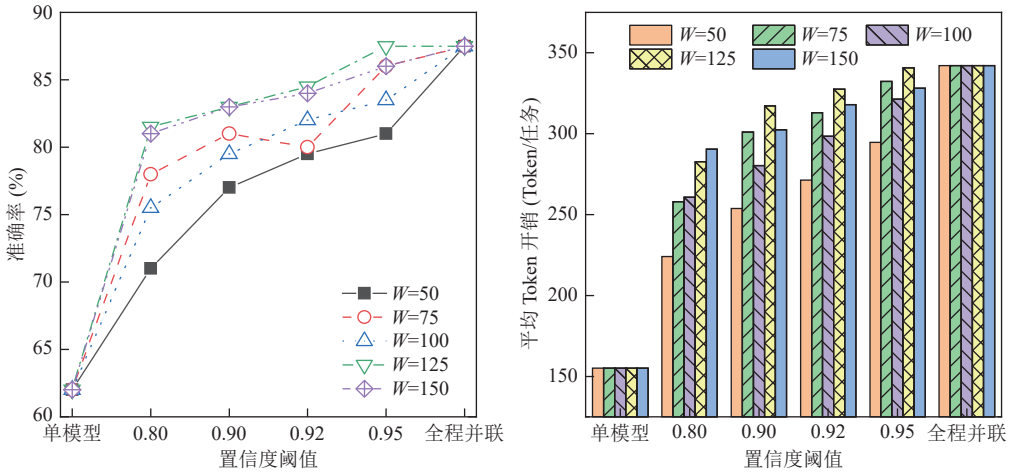
UniTe[10]: 一种基于 Top-K 截断的概率值向量维度的 Token 级并联协作方法. 该方法采用全程 Token 级并联协作策略, 本文沿用其原文中的参数设置, 并将 K 值设定为 10.

Para-N: 该方法采用固定协作长度策略, 即在生成前 N 个 Token 时进行 Token 级并联协作推理, 后续 Token 则由基模型独立完成. 该方法为固定协作次数的消融实验基线, 用于评估 CobCo 中置信度评估算法的有效性.

CE-CoLLM[21]: 一种采用基模型置信度的路由方法, 其置信度仅由基模型自身输出决定, 具体采用最大概率 Token 的概率作为度量. 为适配本文场景, 本文将其思想扩展至 Token 级并联协作中, 当基模型置信度高于阈值时执行独立推理, 否则触发并联协作.

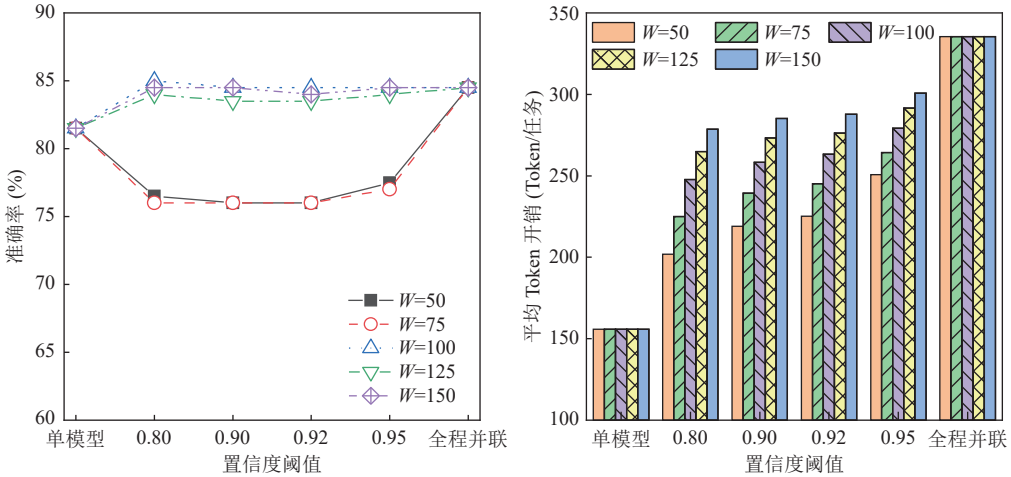
3.2 参数影响分析

本节分析了不同观察窗口大小 W 和阈值 φ 这两个关键参数对推理性能的影响. 实验主要考察了 CobCo 在两类典型场景下的表现: 基模型与协作模型性能差异较大, 以及性能相近. 实验旨在揭示模型性能差异对 CobCo 的普适影响规律, 因此每类场景仅使用一种模型组合, 结果见图 4 和图 5.



(a) 阈值调整对推理准确率的影响　　(b) 阈值调整对应的 Token 开销分析

图 4　基模型与协作模型性能差异较大场景中, 在不同参数下的推理准确率和 Token 开销变化



(a) 阈值调整对推理准确率的影响　　(b) 阈值调整对应的 Token 开销分析

图 5　基模型与协作模型性能相近场景中, 在不同参数下的推理准确率和 Token 开销变化

图 4 为基模型与协作模型性能差异较大情形下的实验结果, 其中基模型为 q1, 协作模型为 q2M. 结果表明, 随着阈值的增加, 基模型的推理准确率呈现先快后慢的增加趋势. 同时, 随着观察窗口 W 增加, 推理准确率亦表现出类似的增长规律. 另一方面, 两个模型整体的平均 Token 开销随阈值或窗口增大而上升. 以上结果表明, 当基模型性能低于协作模型时, 过小的阈值或窗口难以显著提升准确率, 而增大阈值或窗口虽可提高准确率, 但增益逐渐饱和且伴随 Token 开销增加, 因此需在准确率与资源消耗间权衡参数选择.

进一步, 本文分析了基模型与协作模型性能相近的情形 (基模型 q2, 协作模型 g4), 实验结果见图 5. 当观察窗口较小时, CobCo 的推理准确率甚至低于单模型, 这可能是由于过小窗口导致基模型过度自信, 过早切换至独立推理, 使协作阶段不足以形成稳定共识, 从而影响后续独立推理的准确性. 同时, 当 $W \geqslant 100$ 时, 不同置信度阈值下的准确率差异较小, 进一步支持了过度自信的推测. 另一方面, 在合理的阈值与窗口设置下, CobCo 在保障推理准确率不降低的同时, 其平均 Token 开销显著低于全程 Token 级并联协作方法. 这说明 CobCo 在性能相近场景中实现了准确率提升与 Token 开销降低的兼顾. 图 5 的结果同样表明, 过大的观察窗口会显著增加 Token 开销, 而准确率提升有限. 综合图 4 与图 5 的实验结果, 本文在后续实验中将观察窗口大小设定为 100. 对于置信度阈值, 两组实验均显示较小阈值对准确率提升作用有限, 而较高阈值则显著增加 Token 开销. 基于准确率与资源消耗之间的权衡, 本文在后续实验中将置信度阈值设定为 0.9.

在上述实验基础上, 本文进一步对惩罚系数进行敏感性分析, 考察其在两种典型场景下对推理准确率与 Token 开销的影响. 惩罚系数取值范围为 1–5, 其余参数保持不变, 实验结果如图 6 所示. 从实验结果可以看出, 在基模型与协作模型性能差异较大的场景 (即 q1&q2M) 中, 随着惩罚系数的增大, 推理准确率呈现先上升后趋于稳定的变化趋势, 并在惩罚系数为 2 时基本达到收敛; 同时, Token 开销亦呈现相似趋势. 这说明适当增大惩罚系数能够抑制基模型过早切换至独立推理, 从而提升推理性能, 但其边际收益有限. 相比之下, 在基模型与协作模型性能相近的场景 (即 q2&g4) 中, 推理准确率整体保持稳定, 对惩罚系数变化不敏感, 而 Token 开销则随惩罚系数的增加呈现单调上升趋势. 这主要是由于惩罚系数增大使得基模型置信度增长更加保守, 延缓其置信度高于阈值的时间, 从而延长协作持续阶段并增加计算开销. 然而, 由于该场景中基模型已具备较强推理能力, 协作持续时间的变化对最终结果影响有限, 因此准确率未出现显著波动.

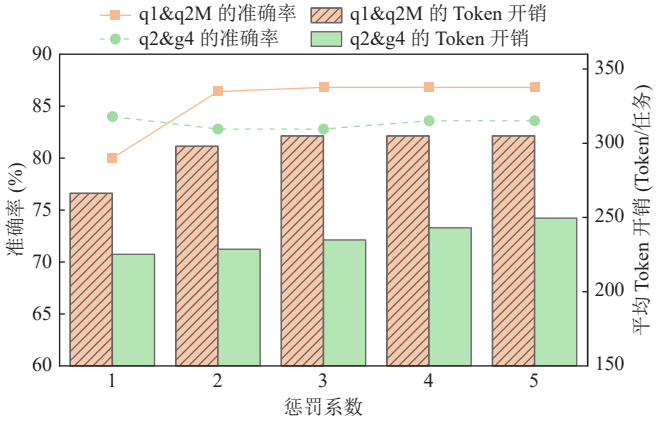


图 6　两种典型场景下, 不同惩罚系数对推理准确率和 Token 开销的影响

综合来看, 惩罚系数在推理准确率与计算开销之间起到关键调节作用: 较小的惩罚系数易导致协作阶段过早终止, 从而影响推理准确性; 而较大的惩罚系数虽有助于稳定性能, 但会显著增加 Token 开销. 该结论从实验层面验证了所提动态协作机制中惩罚设计的必要性与有效性. 因此, 本文在综合权衡后将惩罚系数设定为 2, 以在推理准确率与计算开销之间取得较优平衡.

3.3　推理准确率分析

本节对比分析了单协作模型与多协作模型场景下, 不同方法在 5 类测试数据集上的平均推理准确率. 实验结

果如表 4 和表 5 所示, 其中 Para-*N*(100) 表示 Para-*N* 方法中参数 *N*=100. 为保持对比的一致性, 本文将 Para-*N* 方法的 *N* 值设为 100, 该取值与本文方法 CobCo 中设定的最小观察窗口大小保持一致. 表 4 为单协作模型场景的实验结果, 其中“q1&q2”表示基模型为 q1、协作模型为 q2, 其余组合同理.

表 4　单协作模型场景下平均推理准确率对比

场景	组合	单模型	DuetNet	UniTe	Para-*N*(100)	CobCo
基模型性能差于协作模型	q1&q2	0.644	**0.808** (+16.4%)	0.800 (+15.6%)	0.768 (+12.4%)	0.792 (+14.8%)
	q1&g4	0.644	0.696 (+5.2%)	**0.792** (+14.8%)	0.668 (+2.4%)	0.668 (+2.4%)
	q1&q2M	0.644	**0.860** (+21.6%)	0.844 (+20.0%)	0.756 (+11.2%)	0.812 (+16.8%)
	L3&g4	0.556	0.752 (+19.6%)	**0.800** (+24.4%)	0.736 (+18.0%)	0.748 (+19.2%)
	L3&q2	0.556	0.812 (+25.6%)	**0.852** (+29.6%)	0.800 (+24.4%)	0.812 (+25.6%)
	平均	0.6088	0.7856 (+17.68%)	**0.8176** (+20.88%)	0.7456 (+13.68%)	0.7664 (+15.76%)
基模型性能优于协作模型	q2&q1	**0.816**	0.808 (−0.8%)	0.800 (−1.6%)	0.752 (−6.4%)	**0.816** (+0.0%)
	g4&q1	0.748	0.696 (−5.2%)	**0.792** (+4.4%)	0.700 (−4.8%)	0.700 (−4.8%)
	q2M&q1	**0.888**	0.860 (−2.8%)	0.844 (−4.4%)	0.876 (−1.2%)	0.868 (−2.0%)
	g4&L3	0.684	0.752 (+6.8%)	**0.800** (+11.6%)	0.740 (+5.6%)	0.736 (+5.2%)
	q2&L3	0.812	0.812 (+0.0%)	**0.852** (+4.0%)	0.832 (+2.0%)	0.820 (+0.8%)
	平均	0.7896	0.7856 (−0.4%)	**0.8176** (+2.8%)	0.7800 (−0.96%)	0.7880 (−0.16%)
基模型与协作模型性能相近	g4&q2	0.748	**0.840** (+9.2%)	0.824 (+7.6%)	0.784 (+3.6%)	**0.840** (+9.2%)
	q2&g4	0.816	0.840 (+2.4%)	0.824 (+0.8%)	**0.848** (+3.2%)	0.840 (+2.4%)
	平均	0.782	**0.840** (+5.8%)	0.824 (+4.2%)	0.816 (+3.4%)	**0.840** (+5.8%)

注: 加粗数据表示该行的最高值, 括号内数据表示相对于基模型独立推理的准确率绝对增幅

表 5　多协作模型场景下平均推理准确率对比

协作规模	组合	单模型	DuetNet	UniTe	Para-*N*(100)	CobCo
双协作模型	q1&q2&g4	0.644	0.812 (+16.8%)	**0.828** (+18.4%)	0.720 (+7.6%)	0.732 (+8.8%)
	q1&g4&q2M	0.644	**0.868** (+22.4%)	0.852 (+20.8%)	0.816 (+17.2%)	0.816 (+17.2%)
	q1&q2&q2M	0.644	**0.868** (+22.4%)	0.832 (+18.8%)	0.744 (+10.0%)	0.792 (+14.8%)
	q2&g4&q2M	0.816	**0.880** (+6.4%)	0.844 (+2.8%)	0.860 (+4.4%)	0.864 (+4.8%)
	g4&q2&q2M	0.748	**0.880** (+13.2%)	0.844 (+9.6%)	0.796 (+4.8%)	**0.880** (+13.2%)
	L3&q2&g4	0.556	**0.848** (+29.2%)	0.704 (+14.8%)	0.832 (+27.6%)	0.832 (+27.6%)
	平均	0.675	**0.859** (+18.43%)	0.817 (+14.23%)	0.795 (+11.93%)	0.819 (+14.43%)
三协作模型	q1&q2&g4&q2M	0.644	**0.868** (+22.4%)	**0.868** (+22.4%)	0.756 (+11.2%)	0.768 (+12.4%)
	q2&q1&g4&q2M	0.816	**0.868** (+5.2%)	**0.868** (+5.2%)	0.852 (+3.6%)	0.852 (+3.6%)
	L3&q2&g4&q2M	0.556	0.860 (+30.4%)	**0.896** (+34.0%)	0.804 (+24.8%)	0.796 (+24.0%)
	平均	0.672	0.865 (+19.33%)	**0.877** (+20.53%)	0.804 (+13.20%)	0.805 (+13.33%)

注: 加粗数据表示该行的最高值, 括号内数据表示相对于基模型独立推理的准确率绝对增幅

基于表 4 所示结果, 主要有以下发现.

(1) 在基模型性能差于协作模型时, 相比于基模型独立推理, CobCo 的平均推理准确率提高了约 15.76%. 与 DuetNet 和 UniTe 这两类采用全程 Token 级并联协作策略的方法相比, CobCo 的推理准确率最高降幅约 5.12%. 此外, 在所有模型组合中, CobCo 的推理准确率均优于 Para-*N*. 其原因在于 Para-*N* 采用固定协作 Token 数策略, 缺乏对基模型置信度的动态评估, 易导致过早独立推理并引发错误.

(2) 当基模型性能优于协作模型时, 多数方法的推理准确率低于基模型独立推理. 这表明当基模型本身性能较强时, 与较弱模型进行 Token 级并联协作反而可能限制其性能表现.

(3) 当基模型与协作模型性能相近时, 相比于基模型独立推理, CobCo 的推理准确率相较独立推理平均提升约 5.8%, 相较 Para-*N* 提升约 2.4%, 且与 DuetNet 和 UniTe 相比能够保持准确率不下降. 这表明 CobCo 的基于置信度的推理调节机制可有效避免性能相近模型间的无效协作干扰.

综上, 当基模型性能未显著优于协作模型时, CobCo 能有效提升推理准确率, 最高提升达 25.6% (组合 L3&q2); 而在基模型性能占优的场景中, 协作收益有限. 与全程 Token 级并联协作方法相比, CobCo 在模型性能相近时基本维持原有准确率, 在基模型性能较差时也仅导致较小的准确率损失. 此外, 由于未考虑基模型置信度, Para-*N* 方法易在置信度不足时提前转入独立推理, 从而导致更为显著的准确率下降, 该结果从侧面验证了置信度评估在抑制准确率损失中的关键作用.

相较于采用全程 Token 级并联协作策略的方法, CobCo 在部分场景下出现的轻微准确率下降, 主要源于以下两方面原因: 一是基模型自身推理能力的上限约束, 当其能力较弱时, 即使在协作阶段获得了有效引导, 后续独立推理仍可能产生偏差; 二是基于置信度的提前退出机制在个别情况下可能产生偏差估计, 导致过早结束协作, 未能充分利用 Token 级并联协作的纠错能力. 例如, 在“q1&q2”组合中, 由于 q1 的独立推理能力较弱 (准确率为 0.644), CobCo 在转入独立推理后更容易受到基模型能力上限的影响, 从而导致性能低于全程 Token 级并联协作方法 (DuetNet 和 UniTe). 相比之下, 在“g4&q2”组合中, 由于 g4 具有较强的推理能力 (准确率为 0.748), 即使在置信度驱动下转入独立推理, 仍能够保持较高的准确率. 上述结果表明, CobCo 的性能在一定程度上依赖于基模型能力以及置信度估计的准确性.

随后, 本文分析了多协作模型场景下各方法的推理准确率, 实验结果如表 5 所示. 表中“q1&q2&g4”表示基模型为 q1, 协作模型为 q2 和 g4, 其余情形同理.

结果显示, 相较于基模型独立推理, CobCo 在不同协作模型数量下均显著提升推理准确率, 且基模型性能越弱, 提升幅度越大. 当协作模型数量增至 2 和 3 时, 相比于单模型独立推理, CobCo 仍分别实现约 14.43% 和 13.33% 的平均准确率提升. 与全程 Token 级并联协作方法 (DuetNet、UniTe) 相比, CobCo 在多协作模型场景下仅引入有限的准确率损失, 平均损失不超过 7.2%, 表明其在降低协作强度的同时仍能保持较为稳健的推理性能. 另外, 相较于 Para-*N* 方法, CobCo 在多协作模型数量场景中取得更高的推理准确率, 最高提升达 8.4% (g4&q2&q2M 组合), 进一步验证了基于置信度动态调节协作过程的有效性.

上述结果表明, 在多协作模型场景下, CobCo 仍能显著提升基模型的推理准确率, 且相较于全程 Token 级并联协作策略仅产生有限的准确率损失. 正如前述实验分析所指出的, 受基模型自身性能上限的制约, 即使在 Token 级并联协作阶段已获得正确的推理路径, 基模型在后续独立推理中仍可能产生错误. 这一根本局限导致 CobCo 存在一定的推理准确率损失. 需要指出的是, 随着协作模型数量增加, CobCo 的推理准确率增益略有下降. 这是由于 Token 级并联协作的推理正确性依赖于所有模型的共同贡献, 随着协作模型数量增加, 该依赖性越强; 因此基模型切换到独立推理后可能难以单独延续该高度依赖协作共识的推理路径, 导致推理准确率有所下降.

3.4 Token 开销分析

如前所述, 尽管 CobCo 在部分场景下会引入一定的推理准确率损失, 但其在置信度充分时及时切换至基模型独立推理, 整体 Token 开销将随之降低. 因此, 本节从平均 Token 开销的角度评估 CobCo 的效率性能. 在与前述实验设置一致的条件下, 分别分析单协作模型与多协作模型场景, 实验结果分别如表 6 和表 7 所示.

表 6 展示了单协作模型场景的平均 Token 开销. 由结果可见, 首先, 在基模型性能弱于协作模型的场景中, CobCo 的 Token 开销较 DuetNet 和 UniTe 分别降低 5.9% 和 4.4%. 这是由于在该场景下基模型性能受限, 置信度提升缓慢, 难以满足高阈值条件, 进一步延缓了独立推理阶段的触发, 因此 Token 开销下降幅度相对有限. 其次, 当基模型性能优于协作模型时, 相较于基模型性能较弱的场景, CobCo 的 Token 开销降幅更为显著. CobCo 的 Token 开销较 DuetNet 和 UniTe 分别平均降低约 12.6% 和 11.3%. 这是由于基模型性能占优, 其置信度快速上升, 促使推理过程更早转入独立阶段, 因此 Token 开销相较于基模型性能较差场景更低. 最后, 在基模型与协作模型性能相近的场景中, CobCo 同样实现了 Token 开销的显著降低, 相较于 DuetNet 方法平均降低 11.4%. 这一效果源于基模型与协作模型性能相近, 使其置信度得以较快提升, 从而促使推理过程及时转入独立推理阶段.

综合来看, 随着基模型与协作模型性能差距缩小甚至反超, CobCo 在降低 Token 开销方面的优势愈发显著. 尤其在性能相近场景中, CobCo 在不降低推理准确率的前提下, Token 开销最高可减少 20.8%. 需要说明的是, 在单

协作模型场景中, 尽管 Para-*N* 方法在 Token 开销上降幅更大, 但由表 4 结果可见, 其同时伴随着较为显著的准确率损失. 相比之下, CobCo 在兼顾准确率的前提下实现了更优的综合性能.

表 6　单协作模型场景下平均 Token 开销对比 (Token/任务)

场景	组合	DuetNet	UniTe	Para-*N*(100)	CobCo
基模型性能差于协作模型	q1&q2	281.74	284.83 (−1.1%)	214.90 (23.7%)	272.46 (3.3%)
	q1&g4	249.89	235.90 (5.6%)	205.68 (17.7%)	224.22 (10.3%)
	q1&q2M	295.86	293.14 (0.9%)	223.78 (24.4%)	269.44 (8.9%)
	L3&g4	253.03	228.38 (9.7%)	214.03 (15.4%)	235.82 (6.8%)
	L3&q2	277.20	295.01 (−6.4%)	218.40 (21.2%)	276.03 (0.4%)
	平均	271.54	267.45 (1.5%)	215.36 (20.7%)	255.59 (5.9%)
基模型性能优于协作模型	q2&q1	281.74	284.83 (−1.1%)	217.68 (22.7%)	214.10 (24.0%)
	g4&q1	249.89	235.90 (5.6%)	199.22 (20.3%)	248.96 (0.4%)
	q2M&q1	295.86	293.14 (0.9%)	234.70 (20.7%)	246.41 (16.7%)
	g4&L3	253.03	228.38 (9.7%)	207.86 (17.9%)	247.98 (2.0%)
	q2&L3	277.20	295.01 (−6.4%)	215.99 (22.1%)	228.86 (17.4%)
	平均	271.54	267.45 (1.5%)	215.09 (20.8%)	237.26 (12.6%)
基模型与协作模型性能相近	g4&q2	292.59	255.73 (12.6%)	210.60 (28.0%)	286.50 (2.1%)
	q2&g4	292.59	255.73 (12.6%)	220.79 (24.5%)	231.80 (20.8%)
	平均	292.59	255.73 (12.6%)	215.70 (26.3%)	259.15 (11.4%)

注: 括号内数据表示相对于DuetNet的平均Token开销相对降低比

表 7　多协作模型场景下平均 Token 开销对比 (Token/任务)

协作规模	组合	DuetNet	UniTe	Para-*N*(100)	CobCo
双协作模型	q1&q2&g4	398.23	391.60 (1.7%)	281.09 (29.4%)	306.65 (23.0%)
	q1&g4&q2M	423.67	386.38 (8.8%)	297.31 (29.4%)	315.58 (25.5%)
	q1&q2&q2M	441.91	433.19 (2.0%)	306.48 (30.6%)	364.66 (17.5%)
	q2&g4&q2M	431.82	427.04 (1.1%)	304.33 (29.5%)	307.85 (28.7%)
	g4&q2&q2M	431.82	427.04 (1.1%)	295.16 (31.6%)	399.10 (7.6%)
	L3&q2&g4	383.75	359.84 (6.2%)	278.75 (27.4%)	291.94 (23.9%)
	平均	418.53	404.18 (3.4%)	293.85 (29.8%)	330.96 (20.9%)
三协作模型	q1&q2&g4&q2M	574.22	546.26 (4.9%)	376.89 (34.4%)	391.73 (31.8%)
	q2&q1&g4&q2M	574.22	546.26 (4.9%)	383.74 (33.2%)	386.57 (32.7%)
	L3&q2&g4&q2M	606.30	584.08 (3.7%)	390.06 (35.7%)	390.33 (35.6%)
	平均	584.92	558.86 (4.5%)	383.57 (34.4%)	389.54 (33.4%)

注: 括号内数据表示相对于DuetNet的平均Token开销相对降低比

表 7 为多协作模型场景下的各方法的平均 Token 开销. 结果表明, CobCo 在不同协作规模下均展现出显著的 Token 开销节省优势.

随着协作模型数量增加, CobCo 在降低 Token 开销方面的优势愈加明显. 在双协作模型场景中, CobCo 相较于 DuetNet 和 UniTe 的 Token 开销分别降低 20.9% 和 18.1%; 在三协作模型场景中, 该降幅进一步扩大, 分别为 33.4% 和 30.3%. 结合表 5 结果可见, CobCo 在多协作场景中的平均推理准确率最大降幅仅为 7.2%. 该结果验证 CobCo 能以较小的准确率损失换取显著的 Token 开销优化, 且该效率优势随协作模型规模扩大而进一步增强.

相较于 Para-*N* 方法, CobCo 虽引入了略高的 Token 开销, 但其推理准确率更优, 体现出以有限资源代价换取性能增益的优势. 需要指出的是, Para-*N* 对参数 *N* 的设定高度敏感, 当 *N* 值较小时, 基模型易过早转入独立推理, 进而引发推理准确率出现明显下降. 相比之下, CobCo 方法通过基于置信度的自适应推理阶段切换机制, 表现出更强的稳健性与适应性.

3.5　推理时延分析

本节分析了 CobCo 方法在推理时延方面的表现. 如图 7 所示, 实验考察了协作模型数量从 0 至 3 递增时的时

延变化. 图中“q1”为基模型独立推理; “q1&q2 (全)”表示基模型 q1 与协作模型 q2 全程 Token 级并联协作推理; “q1&q2”对应基模型 q1 与协作模型 q2 在 CobCo 下的推理时延, 其余同理.

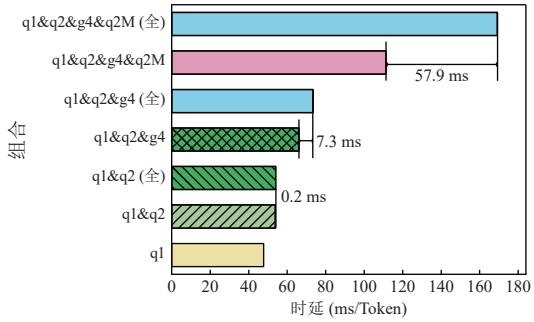


图 7 不同协作模型数量下的推理时延对比

根据图 7 结果, 基模型 q1 独立推理时的时延最低, 约为 47.8 ms/Token. 主要原因在于其推理过程完全在本地完成, 避免了模型间协作带来的通信与同步开销. 随着协作模型数量的增加, CobCo 与全程 Token 级并联协作方法的推理时延均显著上升, 这验证了本文所指出的时延“短板效应”. 在 Token 级并联协作机制下, 整体生成速度受限于协作组合中响应最慢的模型; 协作规模越大, 包含低速模型的概率越高, 从而进一步抬升平均推理时延.

在不同协作规模下, CobCo 的推理时延始终低于全程 Token 级并联协作方法, 且随着协作模型数量增加, 二者的时延差距由 0.2 ms 扩大至 57.9 ms. 相较于全程 Token 级并联协作策略, CobCo 在单 Token 推理时延上最高可降低约 34%. 这一优势源于其在基模型置信度充分时及时切换至独立推理, 从而避免长期受限于低速协作模型, 有效缓解 Token 级并联协作中的时延短板效应. 上述结果从推理时延维度进一步验证了 CobCo 的有效性.

3.6 非全程 Token 级并联协作方法对比分析

考虑到模型组合数量呈指数增长, 本文选取若干具有代表性的组合开展了对比实验, 选定的组合覆盖不同能力差异及协作规模, 实验结果如图 8 所示, 其中 CE-CoLLM(0.8) 表示其阈值为 0.8, 其余类推.

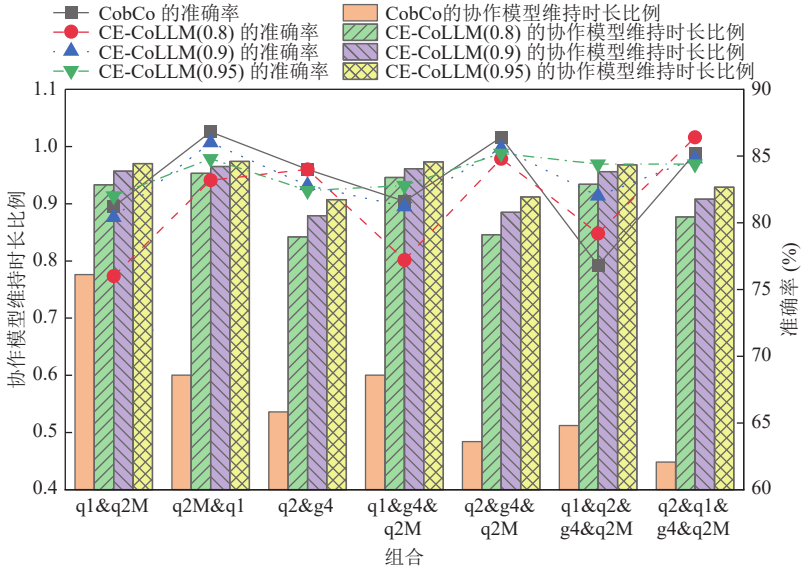


图 8 非全程 Token 级并联协作方法的表现对比

在开销评估方面, 本节采用协作模型维持时长比例进行衡量. 这是由于 CobCo 采用阶段性协作策略, 其输入

Token 的结构相对稳定; 相比之下, CE-CoLLM 采用动态路由机制, 协作模型在推理过程中按需参与, 可能导致输入上下文被重复计算, 从而使 Token 开销具有路径依赖性, 难以进行统一且公平的度量. 因此, 本文引入协作模型维持时长比例作为补充指标, 用于衡量协作模型在任务执行过程中的资源占用情况. 具体而言, 若协作模型在第 N_{last} 个 Token 处最后一次参与推理, 而任务总 Token 数量为 N, 则其维持时长比例定义为 N_{last}/N. 协作模型维持时长比例反映了协作模型在推理过程中需保持可用状态的时间范围. 该指标越高, 表示协作模型需长时间在线以支持潜在调用, 从而增加系统资源占用与调度开销; 反之则说明其可更早释放资源, 有利于提升系统效率.

图 8 实验结果表明, 在推理准确率方面, CobCo 方法整体优于 CE-CoLLM, 仅在 q1&q2&g4&q2M 组合下略低. 其原因在于, 当基模型性能较弱且长期依赖协作时, 协作阶段形成的推理路径更多由群体主导, 基模型在退出后难以独立延续该推理过程, 从而影响最终结果. 而 CE-CoLLM 以基模型为主导, 仅在低置信度时引入协作, 使其更易保持一致的推理路径, 因此在该特定场景下表现略优.

实验结果表明, 在协作模型维持时长比例方面, CobCo 表现出显著优势. 如图 8 所示, CobCo 的协作模型维持时长比例明显低于 CE-CoLLM, 最高由 0.929 降低至 0.448 (组合 q2&q1&g4&q2M), 所有组合的平均降幅约 40%. 这表明, 相较于 CE-CoLLM, CobCo 能够显著缩短协作模型的活跃区间, 使其更早退出当前任务并释放计算资源, 从而降低系统资源占用并提升资源利用效率. 这一差异主要源于两者机制的不同. CobCo 采用阶段性协作策略, 当基模型置信度达到阈值后, 协作模型即可退出; 而 CE-CoLLM 由于采用动态路由机制, 为支持后续可能的模型切换, 其协作模型通常需要在更长时间范围内保持可用状态, 因此整体维持时长比例较高.

进一步观察不同模型组合可以发现, CobCo 的维持时长比例随基模型能力变化呈现出良好的自适应特性. 当基模型性能较弱时, 协作阶段持续时间较长, 因此维持时长比例相对较高; 而当基模型性能较强时, 其置信度能够更快达到阈值, 协作模型可提前退出, 从而显著降低资源占用. 这一结果从侧面验证了本文所提置信度驱动动态退出机制的有效性. 综上, CobCo 能够在保证推理准确率的前提下减少协作模型的参与时长, 体现出了更优的性能与开销权衡.

4 总 结

Token 级并联协作推理作为一种高效的多模型协作范式, 已在模型互联网场景中受到广泛关注. 然而, 现有主流方法多采用全程 Token 级并联协作策略, 导致 Token 开销与推理时延显著增加. 为此, 本文提出一种基于置信度的协作推理方法 CobCo. 该方法通过基于置信度的阶段切换机制, 并结合 Beta 分布对基模型置信度进行实时评估, 实现了在置信度超过预设阈值时由 Token 级并联协作推理切换至独立推理, 从而降低 Token 开销和推理时延. 实验结果表明, CobCo 不仅显著提升了基模型在模型互联网环境下的推理准确率, 而且在与全程 Token 级并联协作方法对比中, 仅以较小的准确率损失为代价, 大幅降低 Token 开销. 特别地, 当基模型与协作模型性能相近时, CobCo 能在保持推理准确率不下降的情况下, 将 Token 开销最高降低约 20.8%. 未来研究将围绕本文实验中揭示的关键现象, 进一步探索更精细化的 Token 级并联协作策略与置信度评估方法, 以提升模型间信息交互效率并增强独立推理阶段决策的准确性.

References

[1] Li ZT, Zeng XY, Wang JH, Xiao Y, Liu ZR, Wu JR, Lai JJ, Huang JJ, Long SQ. AI-model network: Concept, current state and future. Journal of Computer Research and Development, 2026, 63(5): 1305–1318 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.202550223]

[2] Liao LL, Tao M, Xie RP, Zhang Y, Yuan HQ. Fine-grained inference task offloading for large language model in industrial edge-cloud collaborative scenarios. Acta Electronica Sinica, 2025, 53(11): 3880–3893 (in Chinese with English abstract). [doi: 10.12263/DZXB.20250411]

[3] Wei W, Yuan X, Du JW, Li YY. A multi-agent collaboration for completing and optimizing bug reports. Journal of Computer Research and Development, 2026, 63(4): 900–917 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.202550693]

[4] Chen Y, Zhao JH, Han HH. A survey on collaborative mechanisms between large and small language models. arXiv:2505.07460, 2025.

[5] Lu JL, Pang ZL, Xiao M, Zhu YC, Xia R, Zhang JJ. Merge, ensemble, and cooperate! A survey on collaborative strategies in the era of large language models. arXiv:2407.06089, 2024.

[6] Chen ZJ, Lu XD, Li JZ, Chen PP, Li ZR, Sun K, Luo YK, Mao QR, Li M, Xiao LK, Yang DQ, Huang X, Ban YK, Sun HL, Yu PS. Harnessing multiple large language models: A survey on LLM ensemble. arXiv:2502.18036, 2025.

[7] 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 with English abstract). http://www.jos.org.cn/1000-9825/7498.htm [doi: 10.13328/j.cnki.jos.007498]

[8] Wang JH, Li ZT, Wu T, Xie ZN, Fan QY, Long SQ. Token-level collaborative reasoning for parallel multi-models. Chinese Journal of Computers, 2025, 48(11): 2579–2593 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2025.02579]

[9] Yu YC, Kuo CC, Ye ZQ, Chang YC, Li YS. Breaking the ceiling of the LLM community by treating token generation as a classification for ensembling. In: Findings of the Association for Computational Linguistics: EMNLP 2024. Miami: ACL, 2024. 1826–1839. [doi: 10.18653/v1/2024.findings-emnlp.99]

[10] Yao YX, Wu H, Liu MY, Luo SC, Han XW, Liu J, Guo ZJ, Song LQ. Determine-then-ensemble: Necessity of top-*k* union for large language model ensembling. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025. 101533–101551.

[11] Hogg MA. Uncertainty-identity theory. Advances in Experimental Social Psychology, 2007, 39: 69–126. [doi: 10.1016/s0065-2601(06)39002-8]

[12] Wang JL, Wang J, Athiwaratkun B, Zhang C, Zou JY. Mixture-of-agents enhances large language model capabilities. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025. 33944–33963.

[13] Yu XL, Xu CM, Chen ZQ, Zhang YD, Lu SL, Yang C, Zhang JN, Yan SC, Hu XB. Visual document understanding and reasoning: A multi-agent collaboration framework with agent-wise adaptive test-time scaling. arXiv:2508.03404, 2025.

[14] Tao MX, Zhao DY, Feng YS. Chain-of-discussion: A multi-model framework for complex evidence-based question answering. In: Proc. of the 31st Int'l Conf. on Computational Linguistics. Abu Dhabi: ACL, 2025. 11070–11085.

[15] Hong SR, Zhuge MC, Chen J, Zheng XW, Cheng YH, Wang JL, Zhang CY, Wang ZL, Yau SKS, Lin ZJ, Zhou LY, Ran CY, Xiao LF, Wu CL, Schmidhuber J. MetaGPT: Meta programming for a multi-agent collaborative framework. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024. 23247–23275.

[16] Xiao ZY, Zhang DX, Wu YG, Xu LL, Wang YJ, Han XW, Fu XJ, Zhong T, Zeng J, Song ML, Chen G. Chain-of-experts: When LLMs meet complex operations research problems. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024.

[17] Qian C, Xie ZH, Wang YF, Liu W, Zhu KL, Xia HC, Dang YF, Du ZY, Chen WZ, Yang C, Liu ZY, Sun MS. Scaling large language model-based multi-agent collaboration. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025. 41488–41505.

[18] Mavromatis C, Karypis P, Karypis G. Pack of LLMs: Model fusion at test-time via perplexity optimization. In: Proc. of the 2024 Conf. on Language Modeling. Philadelphia: OpenReview.net, 2024. 1–21.

[19] Jin LF, Peng BL, Song LF, Mi HT, Tian Y, Yu D. Collaborative decoding of critical tokens for boosting factuality of large language models. arXiv:2402.17982, 2024.

[20] Li TL, Liu Q, Pang TY, Du C, Guo Q, Liu Y, Lin M. Purifying large language models by ensembling a small language model. arXiv:2402.14845, 2024.

[21] Jin HP, Wu YZ. CE-CoLLM: Efficient and adaptive large language models through cloud-edge collaboration. In: Proc. of the 2025 IEEE Int'l Conf. on Web Services. Helsinki: IEEE, 2025. 316–323. [doi: 10.1109/ICWS67624.2025.00046]

[22] Qwen Team. Introducing Qwen1.5. 2024. https://qwen.ai/blog?id=qwen1.5

[23] Yang A, Yang BS, Hui BY, *et al*. Qwen2 technical report. arXiv:2407.10671, 2024.

[24] Zeng AH, Xu B, Wang BW, *et al*. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. arXiv:2406.12793, 2024.

[25] Grattafiori A, Dubey A, Jauhri A, *et al*. The Llama 3 herd of models. arXiv:2407.21783, 2024.

[26] Lin S, Hilton J, Evans O. TruthfulQA: Measuring how models mimic human falsehoods. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022. 3214–3252. [doi: 10.18653/v1/2022.acl-long.229]

[27] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. Measuring massive multitask language understanding. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021. 1–27.

[28] Du YL, Li S, Torralba A, Tenenbaum JB, Mordatch I. Improving factuality and reasoning in language models through multiagent debate.

In: Proc. of the 41st Int'l Conf. on Machine Learning. Vienna: OpenReview.net, 2024. 11733–11763.

[29] Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, Plappert M, Tworek J, Hilton J, Nakano R, Hesse C, Schulman J. Training verifiers to solve math word problems. arXiv:2110.14168, 2021.

[30] Clark C, Lee K, Chang MW, Kwiatkowski T, Collins M, Toutanova K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: ACL, 2019. 2924–2936. [doi: 10.18653/v1/N19-1300]

附中文参考文献

[1] 李哲涛, 曾曦玉, 王建辉, 肖勇, 刘忠仁, 吴俊儒, 赖俊杰, 黄纪俊, 龙赛琴. 模型互联网: 概念、现状和未来. 计算机研究与发展, 2026, 63(5): 1305–1318. [doi: 10.7544/issn1000-1239.202550223]

[2] 廖玲玲, 陶铭, 谢仁平, 张引, 袁华强. 面向工业场景的边-云协同大语言模型细粒度推理任务卸载. 电子学报, 2025, 53(11): 3880–3893. [doi: 10.12263/DZXB.20250411]

[3] 魏威, 苑兴, 杜军威, 李玉莹. 多代理协作实现缺陷报告补全和优化. 计算机研究与发展, 2026, 63(4): 900–917. [doi: 10.7544/issn1000-1239.202550693]

[7] 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 [doi: 10.13328/j.cnki.jos.007498]

[8] 王建辉, 李哲涛, 伍涛, 谢展楠, 樊乾意, 龙赛琴. Token 级多模型并联协作推理. 计算机学报, 2025, 48(11): 2579–2593. [doi: 10.11897/SP.J.1016.2025.02579]

作者简介

王建辉, 博士生, 主要研究领域为边缘计算, 人工智能.

李哲涛, 博士, 教授, 博士生导师, 主要研究领域为计算机网络, 人工智能和安全.

刘忠仁, 博士生, 主要研究领域为大模型推理, 人工智能.

肖勇, 博士生, 主要研究领域为隐私保护, 人工智能.