Token级多模型并联协作推理

王建辉 李哲涛 伍 涛 谢展楠 樊乾意 龙赛琴
(暨南大学信息科学技术学院 广州 510632)

摘 要 推理准确率作为大模型的核心评估指标,对模型的实际应用效果和用户体验具有重要影响。多模型协作推理是提升推理准确率的有效途径之一,其主要分为全回复级协作和Token级协作。Token级协作相比全回复级协作在Token开销和时间成本方面具有显著优势。然而,现有Token级协作方法存在低置信度Token噪声过滤不足以及在聚合过程中平等化模型贡献等问题。为此,本文设计了一种新型Token级模型并联协作推理架构——DuetNet。该架构通过汇聚多个模型的推理共识以降低选择错误推理路径的可能性,从而提高推理准确率。具体而言,在每个推理步骤中,DuetNet首先应用联合截断策略,以减少引入低置信度Token噪声;随后,在聚合过程中,通过聚合逻辑值向量计算每个候选Token的累计逻辑分数,以降低置信度损失;最后,通过Top-T随机采样算法选择下一个Token。实验结果表明,DuetNet框架下的多模型并联协作在推理准确率方面优于现有方法。在双模型并联协作时,DuetNet的平均推理准确率相对于其他方法提高了1.88%～38.50%,并且在推理过程中需要对齐的Token数量减少了80%以上。在三模型和四模型并联协作场景中,DuetNet同样显示出较好的推理准确率提升,相对于其他方法提高了1.21%～40.34%。

关键词 大模型;推理准确率;多模型协作;推理共识;Token级模型并联协作
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.02579

Token-Level Collaborative Reasoning for Parallel Multi-Models

WANG Jian-Hui LI Zhe-Tao WU Tao XIE Zhan-Nan FAN Qian-Yi LONG Sai-Qin
(College of Information Science and Technology, Jinan University, Guangzhou 510632)

Abstract As a core metric for large models, inference accuracy critically influences their practical performance and the user experience. Multi-model collaborative reasoning is one of the effective ways to improve inference accuracy, which is mainly divided into full-response-level collaboration and token-level collaboration. Token-level collaboration has significant advantages over full-response-level collaboration in terms of token overhead and time cost. Nevertheless, existing token-level collaboration methods face issues, such as inadequate filtering of low-confidence token noise and equalizing model contributions during the aggregation process. To address these challenges, this paper designs a novel token-level model parallel collaboration inference architecture—DuetNet. This framework enhances inference accuracy by aggregating the inference consensus from multiple models to reduce the likelihood of selecting erroneous inference paths. Specifically, in each reasoning step, DuetNet first applies a joint truncation strategy to mitigate the introduction of low-confidence token noise. Subsequently, during the aggregation process, it calculates the cumulative logit scores of each candidate token by aggregating the logit value vectors, thereby

收稿日期:2025-03-17;在线发布日期:2025-07-11。本课题得到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)资助. **王建辉**,博士研究生,主要研究领域为边缘计算与人工智能。E-mail:tranfer98@foxmail.com。**李哲涛(通信作者)**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为计算机网络、人工智能和安全。Email:liztchina@hotmail.com。**伍 涛**,硕士研究生,主要研究领域为物联网和人工智能。**谢展楠**,硕士研究生,主要研究方向为多智能体协作机制。**樊乾意**,硕士研究生,主要研究领域为人工智能和安全。**龙赛琴**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为人工智能、云计算和边缘计算。

reducing confidence loss. Finally, the next token is selected using a Top-T stochastic sampling algorithm. Experimental results indicate that multi-model parallel collaboration under the DuetNet framework outperforms existing methods in terms of inference accuracy. In the dual-model parallel collaboration scenario, DuetNet improved average inference accuracy by 1.88% to 38.50% compared to other methods, while the number of tokens requiring alignment during inference was reduced by over 80%. In both tri-model and quad-model parallel collaboration scenarios, DuetNet demonstrates consistent inference accuracy improvements, achieving performance gains of 1.21% to 40.34% over comparative methods.

Keywords large models; inference accuracy; multi-model collaboration; reasoning consensus; Token-level model parallel collaboration

1 引 言

大模型作为人工智能革命的核心突破,正在重塑人机交互方式和技术发展路径。自 ChatGPT 发布以来,其强大的通用能力和广泛应用价值使大模型成为了数字化转型的关键力量和国际科技竞争的焦点[1-3]。尽管大模型显著提升了人工智能的推理能力,但现有模型在推理准确率方面仍存在明显不足。这一局限性促使学术界与工业界持续探索优化推理性能的新方法[4-7]。当前,提升模型推理准确率的技术路径主要归纳为三类:训练规模更大或推理效果更佳的大模型、探究模型内部思维模式和设计多模型协作策略。

训练规模更大或推理效果更佳的大模型是提升模型推理准确率的技术路径之一,该路径的实现途径包括训练参数量更大的大模型[8-9]、优化数据集和训练策略[10-11]以及探索新的模型架构[12-13]。以 OpenAI 系列模型为例,GPT-1 的参数量仅为 1.1 亿,而性能显著提升的 GPT-3.5 其参数量达到 1750 亿,更高性能的 GPT-4 据推测可能具备万亿参数规模。Llama 3、Qwen 2.5 和 Mistral 等一系列基于 Transformer 架构的大语言模型主要通过优化训练数据和训练策略来提升模型推理性能。2024 年 9 月,基准测试平台 Chatbot Arena 公布的大模型盲测榜单结果显示,Qwen2.5-72B-Instruct 位列全球前十,这一结果验证了优化数据集和训练策略的有效性[14]。此外,由于 Transformer 架构在处理较长文本时存在局限性,2023 年 12 月 Mamba 架构被提出并旨在替代 Transformer 架构的主导地位[12],然而该架构的有效性目前仍处于探索阶段。

探索模型内部思维模式是近年来迅速发展的重要研究方向,研究者通过探索将人类的慢思考特性赋能给大模型从而提高模型的推理准确率。2022 年 Google 提出思维链(Chain-of-Thought,CoT)策略,该策略显著增强了大模型在决策过程中的表现[15]。CoT 策略的有效性促使研究人员探索其他思维方式以优化模型表现,包括思维链自洽性[16]、思维树[17]和思维图[18]。OpenAI 的 o1 模型将思维链内置,模型在训练过程中系统性地掌握了多步骤推理能力,从而使得 o1 的推理准确率显著提升[19]。深度求索公司开发的 DeepSeek-R1 模型基于强化学习框架构建了一个自主思维链系统,其在问题解决和创造性任务中展现出显著的智能水平[20]。

上述两类技术路径实质上都是专注于优化单一模型的推理性能并且需要投入高昂的成本重新训练一个大模型。相比之下,多模型协作的技术路径则强调集合多个现有的模型优化推理性能[21-22]。

对于多模型协作,全回复级协作是当前主流方案,旨在通过多个模型对同一问题的完整回答进行协作推理[23-25]。如图 1(a)所示,在全回复级协作的架构下,参与协作的大模型分别根据提示词输出问题的全部回复,然后通过一种协作机制共同优化最终的输出,这类全回复级协作机制包括辩论[26-27]、辩论与反思组合[28]以及讨论[29]等。然而,全回复级协作往往需要多轮协作才会输出最终结果,这导致该方式面临高昂的 Token 开销和时间成本。与全回复级协作相比,Token 级协作只需一轮完整响应即可完成,其推理开销更低。如图 1(b)所示,该方法在每个生成步骤中聚合各模型产生的候选 Token,实现更高效的协作推理[30-33]。

尽管 Token 级协作展现了良好的性能,但现有方法仍面临两个主要问题:(1)现有 Token 级并联协作方法在聚合过程中无法有效过滤低置信度

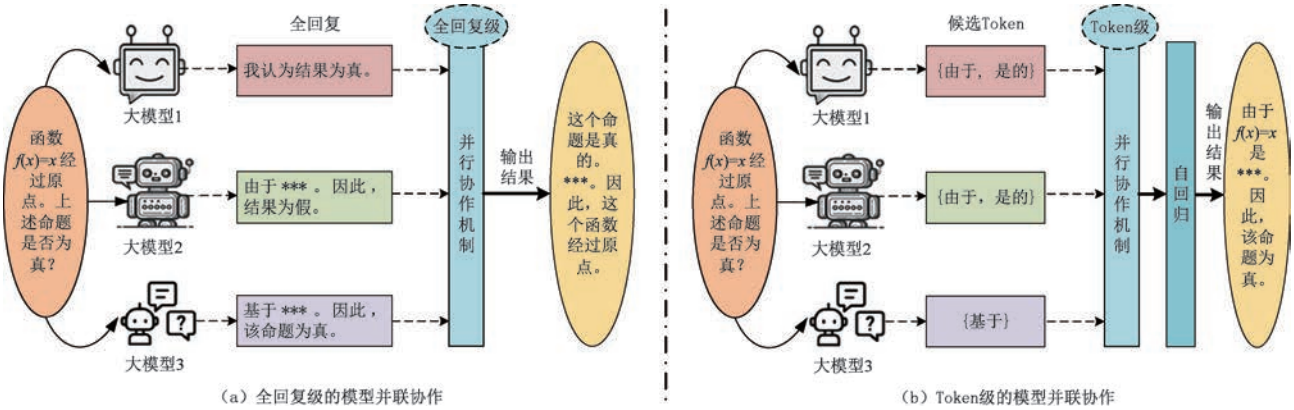

图 1　全回复级协作和 Token 级协作

Token 噪声,影响模型并联协作的推理准确率。具体而言,现有方法中的对齐策略包括全量对齐[30-32]和 Top-K 截断对齐[33]。全量对齐在每个生成步骤中对整个词汇表的概率进行对齐,但易引入低置信度 Token 噪声且计算开销较高。而固定 K 值的 Top-K 截断对齐策略虽然提高了计算效率,但难以适应不同样本的动态概率分布。(2)现有方法[30-33]主要采用归一化后的概率值向量进行等权重聚合,即在聚合时平等化每个模型的贡献,该策略忽视了不同模型预测结果的可靠性差异。例如,当性能优异的模型 A 与欠佳的模型 B 对同一 Token 均给出高置信度预测时,后者可能仅反映随机噪声或局部最优解。因此,这种等权重聚合方式会平等对待各模型的贡献度,不仅可能导致概率空间语义失真,还会均等放大各模型的固有偏差。

综观现有大模型推理准确率优化策略,优化单模型性能因需进行模型重训练而导致较高的实现门槛与成本。另一方面,在多模型协作中,全回复级协作的时间成本和 Token 开销高昂,而 Token 级协作方法仍未得到充分探索。

鉴于此,探索高效的 Token 级模型协作策略具有重要的研究价值。随着开源大模型生态的快速发展,模型私有化部署逐渐成为重要趋势。本地部署不仅能提供更强的隐私保护能力,还能为用户提供细粒度的模型控制权限。然而,受限于终端设备的计算资源,本地部署的大模型往往面临性能瓶颈,且难以支持多模型并行部署。因此,如何高效提升本地大模型的推理准确率成为当前亟待解决的关键问题。受现有多模型协作实践和互联网的“连接即服务”与共享理念的启发,分布式模型可视为网络中的智能节点。未来,在这一框架下,互联的模型之间能够实现协同推理,从而在保持本地化部署优势的同时提升整体推理效能。因此,本文设计了一种 Token 级多模型并联协作推理方法,旨在评估其在未来由大量本地大模型构成的互联系统中的协作潜力。本文的主要贡献归纳如下:

(1) 本文设计了一种 Token 级模型并联协作推理架构——DuetNet。在该架构下,模型可以实现 Token 级并联协作,共同解决同一难题。同时,本文还设计了一种 Token 级多模型并联协作推理算法,该算法通过汇聚多个模型的推理共识以降低选择错误推理路径的可能性,从而提高推理准确率。

(2) 针对现有 Token 级协作方法存在的局限性,本文提出以下两方面的改进:首先,本文设计了一种联合 Top-K 截断与 Top-P 截断的截断策略。该策略综合了 Top-K 方法在局部稳定性方面的优势与 Top-P 方法在全局适应性方面的特点,不仅显著降低了需要聚合的 Token 数量,还实现了两种策略的优势互补。通过过滤低置信度的候选 Token,可有效降低聚合过程中的噪声。其次,本文采用归一化前的逻辑值向量进行聚合操作,这一设计避免了传统归一化处理导致的置信度损失,从而有效缓解了概率值向量平均加权带来的负面影响。

(3) 实验结果表明本文设计的 DuetNet 可有效提高推理准确率。具体而言,与单模型推理的平均准确率相比,DuetNet 中双模型并联的推理准确率提高了 21.44%,且随着并联模型数量的增加,准确率的提升幅度也更大。此外,实验结果表明 DuetNet 的推理准确率高于其他 Token 级协作方法,同时聚合开销更小。在双模型并联的情况下,DuetNet 的平均推理准确率相比其他方法提升了 1.88%至 38.50%,并且在并联推理过程中需要对齐的 Token 数量减少了 80%以上。最后,在 DuetNet 框架下,双模型并联协作推理的延迟仅比单模

型推理多约 2 毫秒，显示出良好的推理效率。

本文的组织结构如下：第 1 节为引言；第 2 节归纳了相关研究工作；第 3 节介绍了 Token 级多模型并联协作推理架构；第 4 节为实验结果分析；第 5 节总结了并联协作的优势以及未来可能面临的挑战；第 6 节为结论。

2 相关工作

在人类社会中，合作、竞争等组织形式能够实现个体难以独立完成的任务。随着大模型技术的演进，学术界与产业界逐渐探索将人类社会的成功的协作经验应用于多模型协作系统中，例如 360 多模型协作框架[34]。当前，多模型协作的范式主要分为基于角色扮演的协作[35]和流程驱动的协作[29]。

基于角色扮演的协作是指让多个大模型智能体扮演不同的角色，从异构视角执行任务，最终聚合不同角色的结果完成一项任务。360 多模型协作中设计了一种基于角色扮演的多模型协作推理框架，其中参与协作的三个大模型分别扮演专家、反思者和总结者[34]。Lu 等人为每个大语言模型分配不同的角色，以对抗大语言模型的同质性[36]。除多模型协同处理单一任务外，利用角色扮演的另一个思路是将问题分解为多个子问题并交给不同角色进行处理。Hong 等人针对软件开发任务设置了产品经理、架构师、项目经理、工程师等角色分别负责不同的子任务[37]。Xiao 等人设计了 11 个角色以解决运筹学问题[38]。基于角色扮演的协作方式的关键在于需要根据特定场景设置相关的角色配置以完成特定的复杂任务。

流程驱动的协作通过模仿人类讨论过程或共识形成机制来增强大模型的推理准确率。多模型讨论框架是典型的流程驱动协作，通常包含以下特征：基于特定的主题构建讨论语境；参与讨论的智能体自主发起讨论；通过预定义的聚合机制生成最终的解决方案或者结论。这类框架的代表工作有基于辩论的协作[26-27]和基于讨论的协作[29]。除了协作机制设计外，部分研究尝试通过为大模型赋予额外属性来优化多模型协作效果。Zhang 等人[28]从社会心理学出发构建了“自负”和“随和”两种性格的智能体并将“辩论”和“反思”两种协作方式进行组合。该工作探索了协作方式和智能体的性格对性能的影响，结果表明协作方式的不同对性能有显著影响，且多模型辩论比单模型反思的效果更好。

上述协作相关研究均是围绕全回复级别的多模型协作机制展开，相比之下，当前针对 Token 级的多模型协作的探索相对有限。Token 级的多模型协作主要通过逐步骤聚合多个模型的输出分布，汇聚模型共识从而优化最终推理结果。Xu 等人[30]提出了 EVA 框架，它使用大语言模型词汇表中的重叠标记作为桥梁，在每对大语言模型之间训练一个投影矩阵实现并联协作推理。Huang 等人提出 DEEPEN[31]，它在集成之前使用锚 Token 将输出概率转换为相对表示，然后在协作后通过梯度下降映射到原始模型的词汇空间。Yu 等人设计了 GaC 框架[32]，其通过统计一个并集词汇表以去除不同模型的词汇表的差异，但是在集成过程中需要进行额外的矩阵乘法。研究工作[30-32]由于需要学习转换参数或采用全量级别的词汇聚合策略，其计算开销较大。针对这些工作的不足，Yao 等人设计了 U-nite 框架[33]，该框架在每个生成步骤中仅对模型输出的 Top-K Token 进行聚合，在显著降低计算复杂度的同时有效提升了推理准确率。

综观上述多模型协作推理的研究，可以发现，全回复级模型协作方法（例如辩论和反思）依赖于多轮完整的响应交互才能实现协作，这导致了较高的计算开销和延迟。相比之下，Token 级并联协作仅需一轮完整的响应即可完成协作，从而实现了更高效的推理过程。然而，现有的 Token 级并联协作方法[30-33]聚焦于词汇表的概率化向量的聚合策略，该策略平等化模型贡献会导致模型的置信度损失，从而影响并联协作的推理准确率。此外，现有 Token 级并联协作方法的对齐策略无法有效过滤低置信度 Token 噪声。因此，针对现有方法的不足，本文设计了一种新的 Token 级模型并联协作推理框架 DuetNet 以优化模型并联协作的推理准确率。

3 Token 级多模型并联协作推理架构

本节详细描述了 Token 级的多模型并联协作推理架构——DuetNet。首先概括了系统模型，然后介绍了并联协作推理的流程，最后设计了一种模型并联协作推理算法实现 Token 级多模型协作推理。

3.1 系统模型

本文考虑的系统模型如图 2 所示。该系统包括一个模型互联平台、若干无本地模型的用户和若干具备单个模型或多个模型的用户。

模型互联平台是整个系统的核心组件之一，负

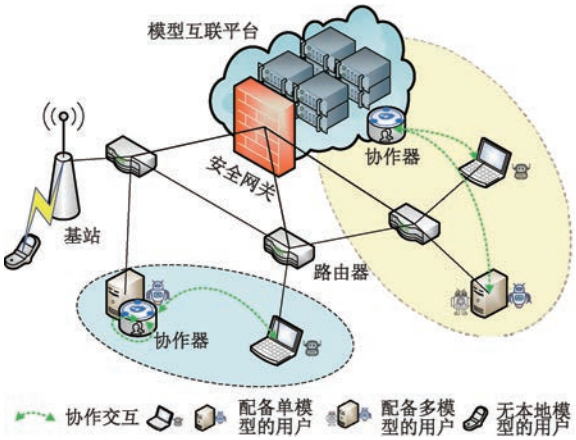

图 2　模型互联系统示意图

责实现模型的接入、管理与协作,确保各个模型之间能够高效、顺畅地相互交流与配合。

用户是模型互联系统的重要组成部分。在该系统中,用户可分为两类:一类为无本地模型用户,他们依赖模型互联平台提供的多种模型以完成特定任务;另一类为具备单个或多个本地模型的用户,他们不仅可以直接使用自身模型,还能通过平台与其他用户的模型进行协作。

协作器是控制模型协作的核心组件之一,多个模型通过协作器共同完成同一个任务。协作器可以部署在发起任务的边缘终端设备上,也可以位于中心平台。多个模型基于协作器协作完成同一任务的流程可概括如下:(1)协作器将任务分发给参与协作的模型;(2)多个模型利用协作器聚合候选 Token 集合,进行自回归生成回复;(3)任务完成后,协作器将结果反馈给相应的用户。

3.2　DuetNet 中并联协作推理流程概述

图 3 系统呈现了 DuetNet 中 Token 级模型并联协作的流程。DuetNet 架构借鉴了音乐二重奏(Duet)的协作理念,通过类似乐器配合的方式,使多个模型能够在 Token 层级实现并联协作,共同解决复杂问题。在图 3 中,每个拼图单元代表一个独立的 Token,而完整的输入序列则通过多个拼图单元的有机组合进行表征。

如图 3 所示,多个模型进行 Token 级并联协作完成同一个任务的流程具体如下:

(1)协作器将问题(输入序列)分发给参与并联协作的多个模型;

(2)每个模型生成下一个 Token 的候选 Token 集合并上传到协作器;

(3)协作器聚合所有参与模型的候选 Token 集合的信息并通过一种协作范式选择出下一个

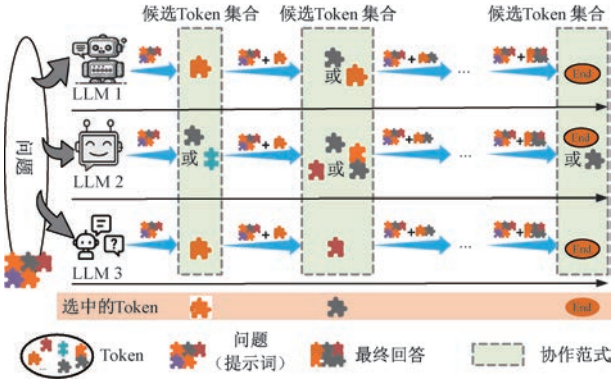

图 3　DuetNet 中 Token 级模型并联协作推理流程

Token,即新 Token;

(4)协作器将该新 Token 同步给每个模型,该新 Token 将作为每个模型的下一个 Token;

(5)重复步骤 1～4,直到下一个 Token 为终止符或者超出最大协作的 Token 数量,则协作结束。

由上述过程可以发现,最终每个模型输出的回答是一致的。

3.3　模型并联协作推理算法

本小节详细介绍了 DuetNet 中的模型并联协作推理算法,重点包括候选 Token 集合的生成、聚合过程以及下一个 Token 的选择方法。

定义 1. 逻辑值. 如图 4 所示,逻辑值向量是大模型在 Softmax 层之前的原始输出结果。作为模型预测过程中的关键中间变量,逻辑值提供了对下一个 Token 生成可能性的量化表征。具体而言,逻辑值可用来比较不同 Token 的相对可能性。在自回归生成的每个生成步骤中,某一 Token 的逻辑值越高,表明模型基于其参数化知识对该 Token 的生成具有更强的统计确定性。

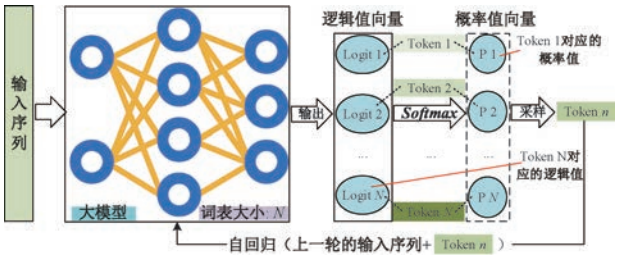

图 4　大模型的自回归生成图示

由图 4 可见,任一模型在每个生成步骤的逻辑值向量经过 Softmax 层后得到的概率值向量都将归一化到 0～1 的取值范围。采用 Softmax 归一化后的概率进行聚合会模糊模型间预测可靠性的差异。这样会导致低质量模型的噪声预测与高质量模型的可靠预测被同等对待,从而扭曲聚合结果的语

义表示并放大模型偏差。对此，本文探索了直接采用逻辑值向量进行聚合的策略。令 N_m 表示大模型 m 的词汇表大小，$\log it_m^n$ 表示第 n 个 Token 对应的逻辑值。那么，对于输入序列 ω，模型 m 生成的逻辑值向量 $\log \boldsymbol{it}_m$ 可以表示为式(1)。

$$\log \boldsymbol{it}_m = G_m(\omega) \tag{1}$$

其中，$\log \boldsymbol{it}_m = \{\log it_m^1, \log it_m^2, \cdots, \log it_m^{N_m}\}$，$G_m(\cdot)$ 表示模型 m 对应的生成函数。为方便后文陈述，$\log \boldsymbol{it}_m$ 表示按照逻辑值大小非升序排序后的逻辑值向量并且 π_m^n 表示 $\log \boldsymbol{it}_m^n$ 对应的 Token。那么，采用 Top-K 与 Top-P 参数分别为 τ_K 和 τ_P 进行联合截断后，所得到的逻辑值向量 $\overrightarrow{\log \boldsymbol{it}}_m$ 可用式(2)表示。

$$\overrightarrow{\log \boldsymbol{it}}_m = \chi(\log it_m, \tau_K, \tau_P) \tag{2}$$

其中，$\chi(\log \boldsymbol{it}_m, T_K, T_P)$ 表示基于参数 τ_K 和 τ_P 从逻辑值向量 $\log \boldsymbol{it}_m$ 中截取满足限制的元素。具体而言，首先从全部候选 Token 中保留概率最高的 τ_K 个 Token，确保候选集覆盖合理的选择范围；其次，基于 Top-P 约束，在 Top-K 截断结果的基础上，仅保留累计概率达到 $\tau_P \in [0,1]$ 的最小 Token 集合。该步骤将排除长尾低概率项，将候选集压缩至模型高置信度的核心支持区间。

假设 M 个模型参与协作，那么在当前生成步骤中所有模型生成的候选 Token 的并集，即联合候选集可表示为式(3)。

$$U = \bigcup_{m=1}^{M} \vec{\pi}_m \tag{3}$$

其中，$\vec{\pi}_m$ 表示 $\overrightarrow{\log \boldsymbol{it}}_m$ 对应的 Token 集合。令 $\boldsymbol{X}_m$ 表示 $\vec{\pi}_m$ 到 U 的映射矩阵，其元素 $X_m(u) \in \boldsymbol{X}_m$ 的取值由式(4)给出。

$$X_m(u) = \begin{cases} \overrightarrow{\log \boldsymbol{it}}_m(u), \text{if } u \in \vec{\pi}_m; \\ 0, \text{else} \end{cases}, \forall u \in U \tag{4}$$

其中，$\overrightarrow{\log \boldsymbol{it}}_m(u)$ 表示模型 m 中 Token u 对应的逻辑值。那么，Token u 的累计逻辑值 C_u 可表示为式(5)。

$$C_u = \sum_{m=1}^{M} X_m(u) \tag{5}$$

算法 1 描述了 DuetNet 中模型并联协作推理算法的具体流程。需要强调的是，DuetNet 框架中的聚合操作直接作用于 Token 语义层面(而非 Token 编号空间)，因此其协作机制与底层分词器实现了有效的解耦。该算法中，每个生成步骤包括 3 个子步骤，即联合截断、聚合和 Top-T 随机采样。首先，每个参与协作的大模型采用联合截断策略生成候选 Token 集合；然后，在聚合步骤计算联合候选集并基于此计算每个候选 Token 的累计逻辑值；最后，对联合候选集进行 Top-T 随机采样。该算法中，Top-T 随机采样的具体过程如下：

(1)从联合候选集选择累加逻辑分数最高的前 T 个 Token 组成待选集合。若 T 值小于联合候选集的大小，则 T 值更新为联合候选集的大小。

(2)基于每个 Token 的逻辑分数大小依概率从待选集合中随机选择出一个元素作为下一个 Token。上述 Top-T 随机采样的过程可表示为式(6)。

$$\hat{u} = Rselect(U, C, T) \tag{6}$$

算法 1. DuetNet 中模型并联协作推理算法

输入：ω；T；$G_m(\cdot), m = 1,2,\cdots,M$；$\tau_K$；$\tau_P$

输出：回答 A

```
A ← ""
DO
    // 步骤 1：联合截断
    FOR 每个大模型 m DO
        logit_m ← G_m(ω)
        ⃗logit_m ← χ(logit_m, τ_K, τ_P)
    END FOR
    // 步骤 2：聚合
    通过式(3)计算候选 Token 集合 U。
    通过式(5)计算 U 对应的累计逻辑值集合 C。
    // 步骤 3：Top-T 随机采样
    û ← Rselect(U, C, T) //依概率随机选择一个 Token
    ω ← ω + û
    A ← A + û
WHILE û ≠ "end" OR Token 数量未超出
RETURN A
```

为便于理解，本文结合图 5 描述 DuetNet 中多模型并联协作推理的生成步骤。

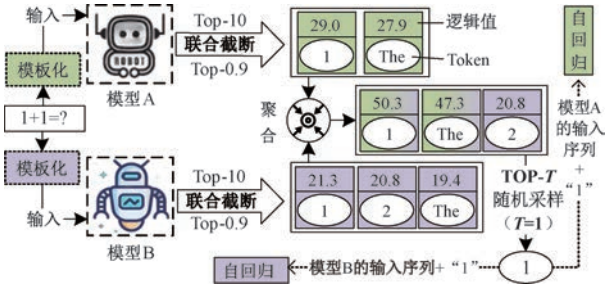

图 5 DuetNet 中模型协作推理的生成步骤示例

如图 5 所示，首先，输入将提示词“1＋1＝?”进行模板化封装，需要强调的是，不同类型模型的聊天模板可能不同。随后，将模板化后的输入序列分别输入到相应的模型并对模型 A 和模型 B 的输出进行联合截断得到候选 Token 集合。紧接着，对两个

模型的候选 Token 集合进行聚合得到联合候选集(即{“1”,“The”,“2”})和对应的累计逻辑值。以 Token “1”为例,其同时出现在两个候选集中,因此其累计逻辑值为两者逻辑值之和,即 29.0+21.3=50.3,其余情形类推。最后,对聚合后的联合候选集执行 Top-T 随机采样,由于图中示例的 T 值为 1,其等效于贪心选择累计逻辑值最高的 Token,故选取 Token“1”。在生成 Token“1”后,将其添加到两个模型的输入序列中进行自回归生成。需注意,自回归过程中均需重复执行前述的联合截断、聚合和 Top-T 随机采样确定下一个 Token。

算法 1 中模型协作推理的生成步骤的复杂度分析。考虑 M 个模型参与协作且每个模型独立部署,模型词汇表规模最大为 $N_{\max}$ 且 Top-K 截断的 K 值为 τ_K。由于 $\tau_K \ll N_{\max}$,因此联合截断步骤的时间复杂度主要来源于 Top-K 截断。由于基于快速排序思想实现的 Top-K 算法的平均时间复杂度为 $O(N_{\max})$,因此,联合截断步骤的时间复杂为 $O(N_{\max})$。联合截断后每个模型的候选 Token 集合最多有 τ_K 个元素,因此,最坏情况下聚合步骤的时间复杂度为 $O(M\tau_K)$。当每个模型生成的候选 Token 集合都不同时,聚合后的联合候选 Token 集合达到最大,即 $M\tau_K$,那么 Top-T 随机采样的时间复杂度为 $O(M\tau_K \log M\tau_K)$。由于 $M\tau_K \ll N_{max}$ 且每个模型独立部署,因此,生成步骤的时间复杂度为 $O(N_{max})$。

4　实验结果分析

4.1　实验设置

如表 1 所示,本文主要选取了四款开源大语言模型进行实验以论证 DuetNet 的有效性。实验在具备 4 块 80G 显存的 A100 的服务器上进行,四款大模型分别部署在不同卡上以模拟模型分布式部署的场景。所有大模型在解码阶段均采用一致的联合截断策略,其中默认设置 $K=\tau_K=10$ 且 $P=\tau_P=0.75$。本文采用表 2 所示的 4 种推理数据集来评估 DuetNet 的性能,包括数学计算、选择题和判断题。本文从每个数据集中随机抽取 50 个题目进行实验测试,并且单模型与多模型并联推理采用的提示词模板一致。

表 1　实验主要选定的 4 款开源大模型

简称	模型名称	发布者	发布时间
$q1$	Qwen1.5-7B-Chat[39]	阿里巴巴集团	2024.02
$q2$	Qwen2.5-7B-Instruct[11]		2024.09
$g4$	glm-4-9B-chat[40]	智谱 AI	2024.06
$L3$	Meta-Llama-3.1-8B-Instruct[10]	Meta	2024.07

表 2　实验选定的 4 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
SimpleMath[26]	简单数学计算题	计算 $a+b\times c+d-e\times f$,其中 a 至 f 为 0～30 的随机整数	{问题}的答案是什么? 请确保在回复的最后陈述你的答案
C-Eval[41]	选择题	中文 AI 大模型评测数据集,涉及 4 个学科大类,52 个学科小类,主要用于评测大模型的知识和逻辑推理能力	{问题}你必须在回答的最后重申你的答案
BoolQ[42]	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
MMLU[43]	选择题	涵盖了美国历史、计算机科学等,难度覆盖高中水平到专家水平的英文数据集	Can you answer the following question as accurately as possible? {**question**} Explain your answer, putting the answer (i.e., A or B or C or D) in the form (X) at the end of your response"

本文对比的方法包括:

(1)SC[16]:一种基于自洽性(Self-consistency,SC)的单模型推理方法。本文使用单个模型在每个生成步骤采用从 Top-1 或 Top-2 Token 中随机选择的策略且每个任务生成 3 条推理路径。

(2)GaC[32]:一种基于全量聚合的概率值向量维度的模型并联协作方法。考虑到大模型输出的 Token 概率分布具有长尾特性(即高概率集中在少数 Token 上),本文采用 Top-50 的候选 Token 进行近似聚合计算。

(3)Unite[33]:一种基于 Top-K 截断的概率值向量维度的模型并联协作方法,基于该工作使用的设置,本文同样将 K 值设定为 10。

(4)DP:一种基于概率值向量的 DuetNet(Du-

etNet based on probability vectors,DP)。该方法是DuetNet的消融方法,其同样采用联合截断策略,与DuetNet的不同之处在于其是基于概率值向量进行聚合。

(5)DK:一种采用Top-K截断的DuetNet(DuetNet with Top-K truncation,DK)。该方法是DuetNet的消融方法,其采用Top-K截断策略生成候选Token(K值设置为10),而聚合阶段和选择下一个Token的策略与DuetNet一致。

4.2 最优 T 值选取

为确定Top-T随机采样子步骤的最优阈值参数T,本节分析了DuetNet中不同双模型并联组合在不同T值下的平均推理准确率。如图6所示,实验设置五种T值方案,其中"随机"是指在每个生成步骤中随机设置T值为1或2,等效$T=1.5$。由结果可见,一方面,不同并联组合的平均推理准确率存在较大的差异;另一方面,平均准确率随T值增大呈现递减趋势。该结果表明在DuetNet架构中,采用累计逻辑分数最高的Token选择策略能较好提升模型并联时的推理性能。

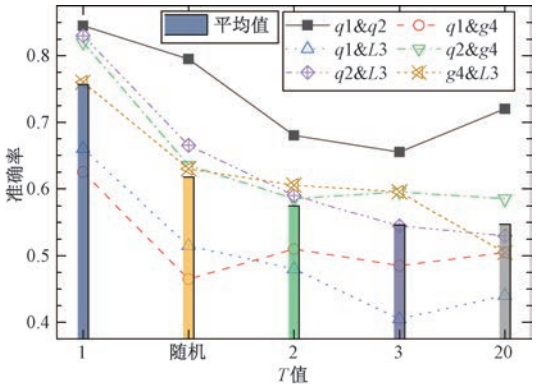

图6 不同取值策略下的平均推理准确率

产生上述结果的一个可能原因是大模型在生成Token的过程中的错误累积效应。一旦选择了错误的Token,后续生成的Token将沿着这一错误路径继续推理,从而导致结果偏差。在DuetNet中,模型并联后在每一个生成步骤中聚合共识,进一步降低了单一模型知识面或性能不足带来的错误推理风险,从而提升了推理准确率。由于模型输出Token的逻辑分数反映了对该选择的信心,逻辑分数越高表示选择的信心越强。因此,贪心选择累计逻辑分数最高的Token在一定程度上可以减少选择到错误方向的可能性。为论证上述猜想,在后续章节中,本文设定T值为1进行了相关实验。

4.3 双模型并联性能评估

本节从推理准确率、推理结果分布、输出相似度和输出Token数量这四个指标评估DuetNet中双模型并联协作推理的性能,指标的详细定义如下:

(1)推理准确率是评估模型并联后推理的准确性,其是衡量推理性能的重要指标。

(2)推理结果分布指单模型的推理结果与模型并联后的推理结果的分布,该指标可更详细地衡量并联推理准确率。

(3)输出相似度指模型并联的输出文本与单模型的输出文本的相似程度。

(4)输出Token数量是将模型并联输出的文本的Token数量与单模型输出的Token数量进行比较以衡量输出长度的差异。

4.3.1 推理准确率分析

表3和表4分别展示了单模型和不同的双模型并联组合的在测试数据集上的推理准确率。表4中的"平均(%, %)"列中的(50.9, 9.0)表示并联组合$q1\&q2$的平均推理性能分别比模型$q1$和$q2$的平均推理性能高50.9%和9.0%,其余情形类推。

表3 单模型的推理准确率(%)

模型	SimpleMath	C-Eval	BoolQ	MMLU	平均
$q1$	50	52	78	44	56
$q2$	72	70	82	86	77.5
$g4$	96	52	84	52	71.5
$L3$	14	48	66	52	45
平均	58	55.5	77.5	58.5	62.5

表4 不同并联组合下的推理准确率(%)

组合	Simple-Math	C-Eval	BoolQ	MMLU	平均
$q1\&q2$	88	78	84	88	84.5↑(50.9, 9.0)
$q1\&g4$	66	50	76	58	62.5(11.6, −12.6)
$q1\&L3$	74	52	76	62	66.0↑(17.9, 46.7)
$q2\&g4$	86	72	84	88	82.5↑(6.5, 15.4)
$q2\&L3$	92	70	84	88	83.5↑(7.7, 85.6)
$g4\&L3$	94	54	80	78	76.5↑(7.0, 70.0)
平均	83.3	62.7	80.7	77.0	75.9

注:↑表示双模型并联协作的推理准确率优于两个单模型的最佳值。

由表3和表4的结果主要得到以下两个发现:

(1)在DuetNet架构下,双模型并联的推理性能明显优于单模型。具体而言,大多数双模型并联组合都优于两个单模型的最佳性能。与单模型相比,双模型并联的平均推理准确率由62.5%增加至75.9%,相对提高21.44%。并联组合$q2\&L3$的性能提升幅度最大,与$q2$和$L3$的平均推理准确率相比(即$(77.5+45)/2=61.25$),$q2\&L3$的推理准确

率提高了 36.3%；而单独与 $L3$ 相比，$q2\&L3$ 的推理准确率提高了 85.6%。

(2)不同的并联组合在推理准确率上表现出显著差异。通常情况下，同系列或性能相近的模型进行并联时，其推理准确率较高。例如，Qwen 1.5 与 Qwen 2.5($q1\&q2$)属于同系列模型，而 Qwen 2.5 与 GLM 4($q2\&g4$)则均为性能优越的模型。从表 4 中可以看出，$q1\&g4$ 和 $q1\&L3$ 的推理准确率相对较低，而 $q1\&q2$ 和 $q2\&g4$ 在各个数据集上的推理性能显著提升，且其准确率均超过了单一模型的最佳推理准确率。这些结果表明，不同的并联组合对推理准确率的影响是显著的，并且选择合适的模型组合能够有效提升整体性推理准确率。

发现(1)说明本文所提出的 DuetNet 在平均维度上可显著提高推理准确率。换言之，从 4 个模型中随机选择 2 个模型进行并联的推理准确率往往优于从中随机选择 1 个模型进行单模型推理的准确率。发现(2)说明不同模型并联显著影响并联的推理准确率，一个可能的原因是模型的版本不同、训练数据不同导致并联推理性能的差异。因此，选择合适的模型进行并联也是一个较为关键的问题。

之后，本文对比了不同框架下的双模型并联推理的平均推理准确率，实验结果如表 5 所示。表中的“$\hat{单}$”列表示两个单模型的最优平均推理准确率。

表 5　平均推理准确率对比(%)

组合	$\hat{单}$	GaC	Unite	DP	DK	DuetNet
$q1\&q2$	77.5	79.0	79.5	78.7	79.0	**84.5**
$q1\&g4$	**71.5**	64.0	62.5	69.5	45.0	62.5
$q1\&L3$	56.0	58.5	62.5	62.5	45.0	**66.0**
$q2\&g4$	77.5	**84.0**	83.5	82.0	54.0	82.5
$q2\&L3$	77.5	78.0	80.5	79.5	51.0	**83.5**
$g4\&L3$	71.5	75.0	**76.5**	75.0	55.0	**76.5**
平均	71.9	73.1	74.1	74.5	54.8	**75.9**

注：带下划线的表示每一行的最高值。

由表 5 的结果可见，DuetNet 优于所有的对比方法，平均推理准确率提高 1.88%～38.50%。GaC、Unite、DP 和 DuetNet 都优于单模型的最优性能，而 DK 的性能远低于其他方法。此外，可以发现 DP 优于 Unite，这说明采用联合截断策略优于采用 Top-K 截断策略，而 DuetNet 优于 DP 说明采用逻辑值向量进行聚合可进一步优化推理准确率。另一方面，DK 的性能较差的原因可能是采用 Top-K 截断和逻辑值向量聚合的策略会引入更大的噪声。

本文进一步分析了 GaC、Unite 和 DuetNet 的每问题所需对齐的 Token 数量，实验结果如图 7 所示。该指标的数值越小，表明计算开销越低，聚合速度越快。由图 7 的结果可见，采用 Top-K 截断策略的 Unite 可以显著降低需对齐的 Token 数量，而采用联合截断策略的 DuetNet 可进一步减少所需对齐的 Token 数量(减少约 80%)。因此，DuetNet 在该指标上显著优于 GaC 和 Unite。

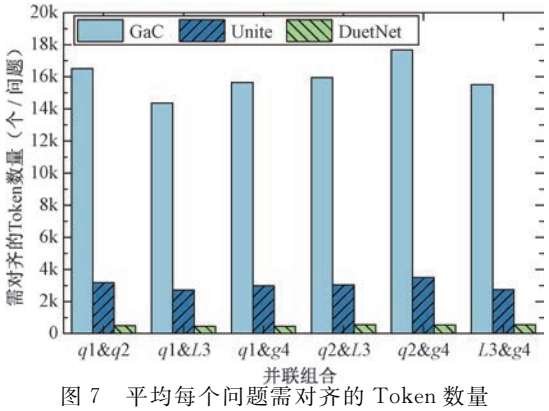

图 7　平均每个问题需对齐的 Token 数量

4.3.2　推理结果分布分析

为分析 DuetNet 优于其他方法的直观原因，本节分析了推理结果分布，结果如图 8 所示。图中“×√√”表示两个单模型的推理结果一错一对但并联后推理正确，“××√”表示两个单模型的推理结果都错但并联后的推理结果正确，其余情况类推。

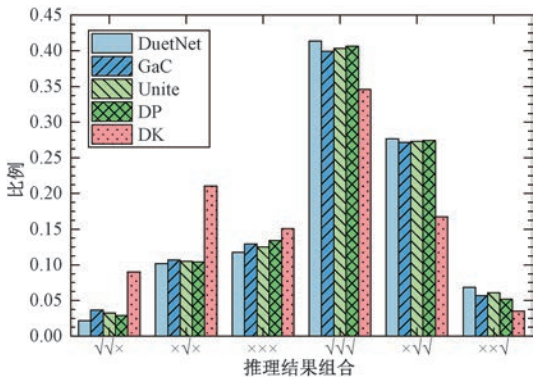

图 8　不同方法的推理结果分布

观察图 8 的结果，主要可发现以下三个现象：

首先，当两个单模型均推理正确时，所有并联方法的推理结果显著偏向正确；当两个单模型的推理结果为一错一对时，除了 DK 方法的推理结果显著偏向错误外，其余方法均显著偏向正确。这些结果表明，当至少一个模型的推理结果正确时，Token 级双模型并联的推理结果更有可能是正确的。

其次，当两个单模型均推理错误时，整体上所有方法的推理结果都倾向于错误，然而，仍然有相当比例的推理结果是正确的。这一结果表明，Token 级

双模型并联具有一定的纠正单模型推理错误的能力，但其纠正效果仍然存在一定的局限性。

最后，与其他方法相比，在相同的组合条件下（例如所有单模型推理均错误），DuetNet 的并联推理结果的错误比例最低，而正确比例最高。由此可见，DuetNet 的推理准确率高于其他方法的直观原因可归纳为“在同等情形下，错的少而对得多”。

4.3.3 输出相似度分析

为深入探讨不同方法的特性，本文分析了双模型并联输出与两个单模型输出在四个测试数据集上的平均余弦相似度的差值，实验结果如图 9 所示。该差值定义为并联输出与模型 1 的相似度减去并联输出与模型 2 的相似度。在相同问题下，双模型并联输出与某个单模型输出的相似度越大，表明并联输出越倾向于该模型，同时也反映出该模型在并联中的主导性更强。因此，相似度差值的绝对值越大，表明并联输出对某一模型的偏向性越明显。例如，在组合 $q1\&q2$ 中，所有方法的差值均小于 0，表明并联输出与模型 $q2$ 的输出更为相似。

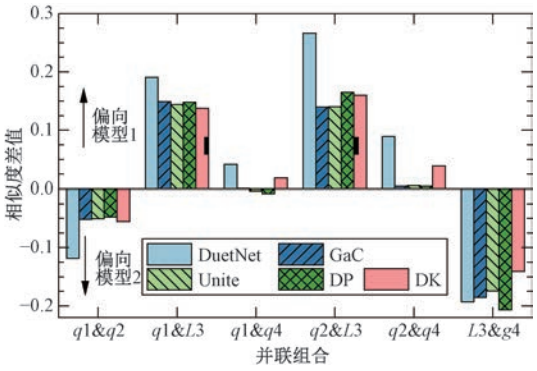

图 9 不同方法下的双模型并联协作的输出相似度差值

从图 9 的结果可以看出，大多数并联组合表现出一定的偏向性，但其程度较小（即相似度差值不超过 0.3）。这表明，在 Token 级模型的并联输出中，通常会出现轻微的“主导”现象。其次，采用逻辑值向量聚合的方法（即 DuetNet、DP 和 DK）的偏向性明显高于采用概率值向量聚合的方法（即 GaC 和 Unite），其中 DuetNet 表现出最强的偏向性。这可能源于采用归一化后的概率值向量聚合会平等化模型的贡献，从而使得相似度差值较小。结合表 5 所示的结果可以发现基于逻辑值向量聚合的 DuetNet 方法的推理准确率高于其他方法，这说明了逻辑值向量的原始尺度差异是有意义的。

4.3.4 输出 Token 数量分析

紧接着，本文探讨了 DuetNet 架构中双模型并联与单一模型在平均每个问题输出的 Token 数量方面的差异，相关实验结果如图 10 所示。由图 10 的结果可见，双模型并联输出的 Token 数量基本介于单模型输出的 Token 数量之间。这说明双模型并联并不会导致输出文本的发散，而是呈现一个更高效的输出质量。需要说明的是，在 DuetNet 中，两个模型以并行方式执行 Token 级协作。这意味着，由于并联模型与单模型在输出文本的 Token 数量上相似，因此两者在延迟方面并无显著差异。

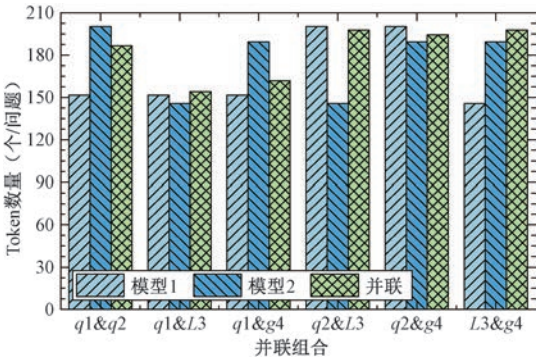

图 10 单模型与并联后输出文本的 Token 数量

为验证上述观点，本文设计了对比实验，以分析 DuetNet 中多模型并联协作推理的延迟，实验结果如图 11 所示。由图中结果可见，在独立推理模式下，模型 $q1$ 和 $q2$ 的单 Token 生成延迟基本相当，分别为 43.94 毫秒和 43.60 毫秒。值得注意的是，当采用 $q1\&q2$ 并联协作时，推理延迟仅比 $q1$ 高 2.51 毫秒，达到 46.45 毫秒。相比之下，模型 $g4$ 的单独推理延迟较高，约为 56 毫秒/Token。由结果可发现推理延迟差异较大的 $q1$ 和 $g4$ 进行并联协作推理时，延迟仅增至 58 毫秒/Token，单 Token 生成延迟相对模型 $g4$ 单独推理仅增加 2 毫秒。

通过对这些结果的深入分析，可以得出两个重要结论。首先，在 DuetNet 的并联协作推理模式下，推理延迟主要受到最慢模型性能的限制。其次，与单一模型推理相比，DuetNet 框架下的并联协作推理生成单个 Token 的延迟的增加幅度非常有限。

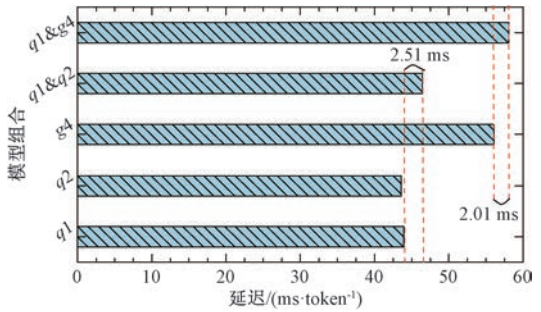

图 11 不同模型组合下的时延比较

这一结果表明,DuetNet 框架下的多模型并联协作推理对用户体验的影响较小。

4.4　同模型与更多模型并联协作性能评估

在本节,本文进行了更多的对比实验分析以论证 DuetNet 的优势。

鉴于 $T=1$ 时并联输出与单模型一致,因此,本文只讨论在次优的随机模式下(即每个生成步骤中 T 随机取 1 或 2)两个同样的模型并联的推理准确率。此外,考虑到同模型并联推理本质上只有一个模型进行推理,因此,本文也引入了单一模型自洽性(即 SC 方法)作为对比基准,其中 $n=3$,即单个模型采样 3 个结果。实验结果如表 6 所示。由结果可见,随机模式下的多数组合的推理准确率高于单模型,SC 方法则优于随机模式。SC 方法的平均推理准确率为 69.4,其低于 DuetNet 中平均推理准确率 75.9(见表 4)。此外,SC 方法需要采样 3 次结果,而双模型并联等效于两个模型采样 1 次结果。因此,SC 方法的 Token 开销高于双模型并联。综上,DuetNet 架构下的双模型并联推理的性能更优。

表 6　随机模式下的平均推理准确率(%)

组合	单模型	随机模式	SC($n=3$)
$q1$&$q1$	56.0	57.0 (+1.0)	63.5 (+7.5)
$q2$&$q2$	77.5	79.5 (+2.0)	86.5 (+9.0)
$g4$&$g4$	71.5	70.0 (−1.5)	74.0 (+2.5)
$L3$&$L3$	45.0	51.5 (+6.5)	53.5 (+8.5)
平均	62.5	64.5	69.4

最后,本文分析了不同方法在更多模型并联时的表现,实验结果如表 7 所示。由结果可见,DuetNet 的平均推理准确率优于其他方法。在三模型的平均指标上,DuetNet 的推理准确率比其他方法提高了 4.51%至 28.84%。在四模型并联时,DuetNet 的准确率提升幅度达到 1.21%至 40.34%。此外,结合表 5 与表 7 的结果可以看出,随着并联模型数量的增加,DuetNet 框架下的模型并联协作推理的平均推理准确率呈现出明显的提升趋势,具体表现为从 75.9 提升至 76.4,再到 83.5。这表明增加并联模型数量可以有效提高 DuetNet 中模型并联的性能。

表 7　多模型并联的平均推理准确率(%)率的均值

组合	均值	GaC	Unite	DP	DK	DuetNet
$q1$&$q2$&$L3$	59.5	81.0	**81.5**	**81.5**	75.0	81.0
$q1$&$q2$&$g4$	68.3	77.5	77.0	**80.0**	71.0	**80.0**
$q1$&$L3$&$g4$	57.5	62.5	60.5	57.5	43.5	**64.0**
$q2$&$g4$&$L3$	64.7	71.5	67.0	68.0	47.5	**80.5**
三模型平均	60.5	73.1	71.5	71.8	59.3	**76.4**
$q1$&$q2$&$g4$&$L3$	62.5	82.5	82.5	78.0	59.5	**83.5**

注:表中的均值表示参与并联的各单模型单独推理时的推理准确率的均值。

4.5　参数实验

本节分析了在联合截断阶段中 Top-K 和 Top-P 的取值对推理准确率的影响。

图 12 展示了不同 Top-P 取值下的平均推理准确率的实验结果。根据图 12 的结果,可以观察到在不同 Top-P 取值下,平均推理准确率表现出先上升后降低的趋势。具体来说,当 P 值增加时,准确率逐渐提高,并在 $P=0.75$ 时达到了最大值 0.757,之后开始下降。这表明在选择 Top-P 时,适当的 P 值对于提升模型的推理准确率是至关重要的。基于实验结果可发现,较小或者较大的 P 值都是不合适的。

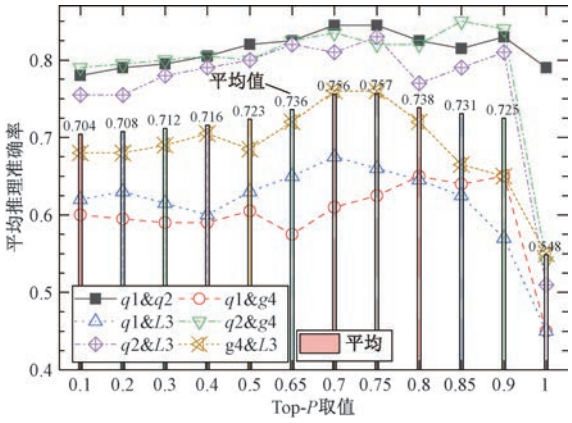

图 12　不同 Top-P 取值下的平均推理准确率
($T=1$, $K=10$)

图 13 展示了不同 Top-K 取值下的平均推理准

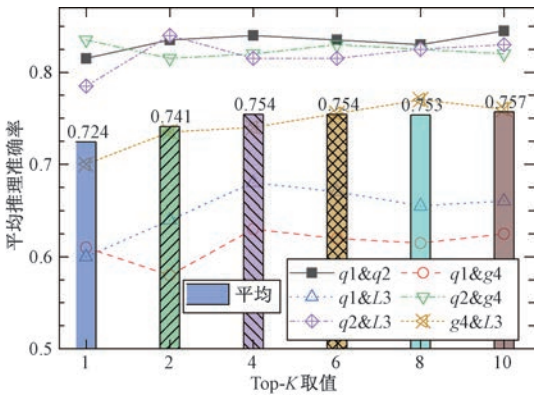

图 13　不同 Top-K 取值下的平均推理准确率
($T=1$, $P=0.75$)

确率的实验结果。根据图 13 的结果，可以观察到，在不同 Top-K 取值下，平均推理准确率呈现出先增加后趋于稳定的趋势。具体来说，随着 K 值的增加，准确率逐渐提高，并在 $K=4$ 时基本收敛。这表明增加 Top-K 的值可以在一定程度上提升模型的性能，但在达到一定值后，进一步增加 K 值对准确率的提升效果变得有限。

5 并联协作优势与未来挑战

5.1 Token 级模型并联协作的优势

类比人类社会中，协作能够相互启发和提升效率从而使得人们高效地完成任务。本文多项实验结果也表明 Token 级双模型并联协作推理相比于单模型推理也有着类似的优势。

首先，多模型并联可以弥补单一模型的不足。在实验结果中，$q1\&q2$ 是同系列模型的不同版本进行并联，$q2\&g4$ 是两个较优的单模型进行并联，这两个组合的并联推理准确率都表现出高于单模型的推理准确率。一个可能的原因是，不同模型受限于各自的训练数据和策略，导致它们在知识水平和推理准确率上存在差异。因此，针对同一推理问题，不同模型可能有不同的理解和思路。合作能够提供多样化的视角，有助于全面分析问题。通过模型并联可以实现知识的互补，从而提高推理准确率。

其次，Token 级模型并联可降低模型推理过程的错误累积，提高推理准确率。大模型推理存在错误累积效应，即某一步推理错误会导致后续的推理都向着错误的方向进行。如图 14 所示，模型 1 在推理错误之后的步骤中推理都偏离了正确答案，而并联协作推理则推理正确。这是由于 Token 级并联协作在生成每个 Token 时汇聚了多个模型的共识，在一定程度上降低单模型的幻觉和知识缺陷问题导致的推理错误。该机制可降低推理过程的错误累积，从而提高推理准确率。本文的实验结果表明多模型并联的推理准确率往往优于随机选择一个单模型，这说明并联协作可降低选择错误推理方向的可能性，从而实现更准确的推理输出。

最后，模型并联可降低用户对模型性能的不确定性。面对多个模型可选的场景，用户往往无法判断一个模型是否全方面优于另一个模型。在心理学中，风险厌恶使得人们存在优先考虑确定性而非不确定性的偏好。如图 14 中所示，针对所测试的问题，若用户只选择一个模型，则平均准确率为 1/2，但选择双模型并联协作，则准确率为 1。这意味着双模型并联协作推理更具确定性，在风险厌恶的心理下，人们对其更具有偏好。本文的实验结果表明多模型并联的推理准确率往往优于随机选择一个单模型，因此用户可以依据经验从候选模型集合中选择两个模型进行并联，从而在一定程度上降低用户对模型推理准确率的不确定性。

5.2 未来挑战

随着模型开源和私有化趋势的加快，如何整合这些私有模型资源，利用现有的模型资源完成复杂的任务成为了一个现实且紧迫的问题。正如本文所设想的一样，在未来，模型将可能互联互通组成一个网络，用户可随时接入网络或共享本地的大模型参与完成复杂任务。该网络可有效整合现有的模型资

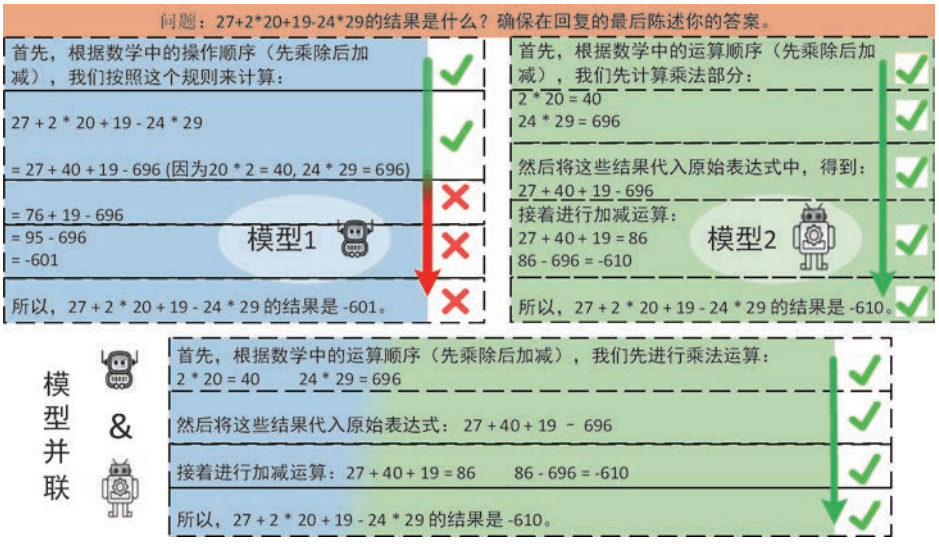

图 14 单模型推理与双模型并联协作推理示例

源,避免模型各自为战,提升模型解决复杂任务的能力。在这种背景下,DuetNet 将会取得进一步发展,但同时也将面临着诸多的挑战。因此,本节总结了未来可能面临的四个挑战。

(1) 探索高效协作

在多模型协作场景中,协作范式是提高推理准确率的关键因素之一,其包括如何针对任务类型选取合适的协作模型、如何根据任务类型选取协作的方式等。本文实验结果已表明不同协作范式的推理准确率有着较大的差异。此外,本文的协作范式依旧存在可改进之处,例如如何在保障推理准确率的同时降低 Token 开销以及如何保障单模型推理正确则并联后也一定推理正确。最后,一些更复杂任务可能需要多种类型大模型进行协作,例如绘画大模型和大语言模型协作完成一项高质量的绘画任务。因此,探索高效的协作方式使得 Token 级的并联协作可完成更复杂任务是未来的挑战之一。

(2) 通信成本优化

由于在 DuetNet 的并联协作中,海量的模型分布式部署在用户终端,两个模型共同完成任务的过程中涉及数据共享。为了改善用户体验和降低网络的负载,多模型协作完成任务的通信成本和延迟应越低越好。因此,如何优化多模型协作过程中的通信成本和通信延迟,提高 Token 级的模型协作的通信效率也是未来亟需解决的挑战之一。

(3) 模型协作激励

DuetNet 的运作依赖于底层模型支持,但出于数据安全和商业利益等考量,模型所有者可能不愿将本地部署的模型接入网络。因此,需要探索如何激励模型所有者将自己的模型接入到网络,激励他们参与模型互联协作;另一方面,还需要探索合理的模型服务定价机制,使得 DuetNet 平台和模型所有者都有所收益,促进 DuetNet 的健康运作。

(4) 安全可信协作

DuetNet 的安全可靠是极为关键的一环,只有实现安全可靠模型互联才能使得用户放心使用其提供的服务。因此,DuetNet 中的安全可信协作问题是未来需要解决的核心挑战之一,例如协作过程中的隐私保护问题、安全通信问题、恶意模型识别问题、外部攻击问题。

6 结　论

为优化大模型的推理准确率,本文设计了一种 Token 级模型并联协作推理架构—DuetNet。在 DuetNet 中,多个大模型可进行更细粒度的 Token 级协作推理以解决同一难题。在 DuetNet 中,多个大模型通过汇聚推理共识降低推理过程的错误累积,从而提高推理准确率。实验结果表明,DuetNet 框架下的模型并联协作推理准确率优于现有方法,并且聚合开销更低。此外,在 DuetNet 框架下,多模型并联协作的平均推理准确率较单模型相对提高了 21.44%以上,而生成单个 Token 的延迟仅增加约 2 毫秒。最后,本文也指出了 DuetNet 在未来将面临若干亟待解决的挑战,包括如何设计更为高效的协作范式、优化通信成本、激励用户参与模型共享以及确保合作过程的安全性和可信性等问题。

参　考　文　献

[1] Kozachek D. Investigating the perception of the future in gpt-3,-3.5 and gpt-4//Proceedings of the 15th Conference on Creativity and Cognition (ACM C&C '23). New York, USA, 2023, 282-287

[2] Zhou Kun, Zhu Yu-Tao, Chen Zhi-Peng, et al. YuLan-Chat: A large language model based on multi-stage course learning. Chinese Journal of Computers, 2025, 48(1): 1-18(in Chinese)
(周昆,朱余韬,陈志朋等. YuLan-Chat:基于多阶段课程学习的大语言模型. 计算机学报,2025,48(1):1-18)

[3] Yao Y, Duan J, Xu K, et al. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 2024, 4(2): 100211

[4] Wang W, Dong L, Cheng H, et al. Augmenting language models with long-term memory//Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). New Orleans, USA, 2023, 14-14

[5] Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv: 2108.07258, 2021

[6] Wang R, Qi J, Chen L, et al. Survey of collaborative inference for edge intelligence. Journal of Computer Research and Development, 2023, 60(2): 398-414 (in Chinese)
(王睿,齐建鹏,陈亮,杨龙. 面向边缘智能的协同推理综述. 计算机研究与发展, 2023, 60(2): 398-414)

[7] Yuan M, Zhang L, et al. Resource-efficient model inference for aiot: a survey. Chinese Journal of Computers, 2024, 47(10): 2247-2273 (in Chinese)
(袁牧,张兰,姚云昊等. 面向智能物联网的资源高效模型推理综述. 计算机学报, 2024, 47(10): 2247-2273)

[8] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv preprint arXiv: 2001.08361, 2020

[9] Bhattacharya P, Prasad V K, Verma A, et al. Demystifying

ChatGPT: An in-depth survey of OpenAI's robust large language models. Archives of Computational Methods in Engineering, 2024, 31, 4557-4600

[10] Dubey A, Jauhri A, Pandey A, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024

[11] Yang A, Yang B, Zhang B, et al. Qwen2. 5 Technical Report. arXiv preprint arXiv:2412.15115, 2024

[12] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023

[13] Liang Y, Wen H, Nie Y, et al. Foundation models for time series analysis: A tutorial and survey//Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. Barcelona, Spain,2024: 6555-6565

[14] Chatbot Arena, https://lmarena.ai/ 2025, 05, 03

[15] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models//Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA, 2022: 24824-24837

[16] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171, 2022

[17] Yao S. Yu D. Zhao J. et al. Tree of thoughts: Deliberate problem solving with large language models//Proceedings of the 37th Conference on Neural Information ProcessingSystems (NeurIPS). New Orleans, USA, 2024: 12345-12360

[18] Jin B. Xie C. Zhang J. et al. Graph chain-of-thought: augmenting large language models by reasoning on graphs//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). Bangkok, Thailand, 2024: 14664-14690

[19] Zhong T, Liu Z, Pan Y, et al. Evaluation of openai o1: Opportunities and challenges of agi. arXiv preprint arXiv: 2409.18486, 2024

[20] Guo D, Yang D, Zhang H, et al. DeepSeek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025

[21] Sun Q, Yin Z, Li X, et al. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. arXiv preprint arXiv:2310.00280, 2023

[22] Feng S, Sorensen T, Liu Y, et al. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. arXiv preprint arXiv: 2406.15951, 2024

[23] Feng S. Shi W. Wang Y. et al. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). Bangkok, Thailand, 2024: 14664-14690

[24] Kenthapadi K. Sameki M. Taly A. Grounding and evaluation for large language models: Practical challenges and lessons learned (survey)//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). Barcelona, Spain, 2024: 6523-6533

[25] Liu L. Zhang D. Li S. et al. Two heads are better than one: Zero-shot cognitive reasoning via multi-LLM knowledge fusion//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM). Boise, USA, 2024: 1462-1472

[26] Du Y, Li S, Torralba A, et al. Improving factuality and reasoning in language models through multiagent debate//Proceedings of the Forty-first International Conference on Machine Learning. Vienna, Austria,2024:11733-11763

[27] Liang T, He Z, Jiao W, et al. Encouraging divergent thinking in large language models through multi-agent debate//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami, USA,2024:17889-17904

[28] Zhang J, Xu X, Zhang N, et al. Exploring collaboration mechanisms for LLM agents: A social psychology view //Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand 2024: 14544-14607

[29] Wang Q, Wang Z, Su Y, et al. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? //Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand,2024: 6106-6131

[30] Xu Y, Lu J, Zhang J. Bridging the gap between different vocabularies for LLM ensemble//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Mexico City, Mexico,2024:7140-7152

[31] Huang Y, Feng X, Li B, et al. Enabling ensemble learning for heterogeneous large language models with deep parallel collaboration. arXiv preprint arXiv:2404.12715, 2024

[32] Yu Y C, Kuo C C, Ye Z, et al. Breaking the ceiling of the LLM community by treating token generation as a classification for ensembling. //Proceedings of the Association for Computational Linguistics: EMNLP. Miami, USA, 2024: 1826-1839

[33] Yao Y, Wu H, Liu M, et al. Determine-then-ensemble: Necessity of Top-k union for large language model ensembling//Proceedings of the Thirteenth International Conference on Learning Representations. Singapore, 2025

[34] Nano Search, https://bot.360.com/ 2024, 12, 24

[35] Shanahan M, McDonell K, Reynolds L. Role play with large language models. Nature, 2023, 623(7987): 493-498

[36] Lu L C, Chen S J, Pai T M, et al. LLM discussion: Enhancing the creativity of large language models via discussion framework and role-play. arXiv preprint arXiv:2405.06373, 2024

[37] Hong S, Zheng X, Chen J, et al. Metagpt: Meta programming for multi-agent collaborative framework. arXiv preprint arXiv:2308.00352, 2023

[38] Xiao Z, Zhang D, Wu Y, et al. Chain-of-experts: When LLMs meet complex operations research problems//Proceedings of the Twelfth International Conference on Learning Representations. Vienna, Austria,2024:Poster

[39] Introducing Qwen1.5, https://qwenlm.github.io/blog/qwen1.5/ 2025, 02, 18

[40] GLM T, Zeng A, Xu B, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793, 2024

[41] Huang Y, Bai Y, Zhu Z, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models// Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems. New Orleans, USA, 2023, 36: 62991-63010

[42] Clark C, Lee K, Chang M W, et al. Boolq: Exploring the surprising difficulty of natural yes/no questions. arXiv preprint arXiv:1905.10044, 2019

[43] Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding. arXiv preprint arXiv: 2009.03300, 2020

WANG Jian-Hui, Ph. D. candidate. His main research interests include edge computing and artificial intelligence.

LI Zhe-Tao, Ph. D., professor, Ph. D. supervisor. His main research interests include computer networks, artificial intelligence and security.

WU Tao, M. S. candidate. His main reaserch interests include Internet of things and artificial intelligence.

XIE Zhan-Nan, M. S. candidate. His main research interests include multi-agent collaboration mechanisms.

FAN Qian-Yi, M. S. candidate. His main research interests include artificial intelligence and security.

LONG Sai-Qin, Ph. D., professor, Ph. D. supervisor. Her main research interests include artificial intelligence, cloud computing and edge computing.

Background

As a core metric for large models, inference accuracy critically influences their practical performance and the user experience. Current approaches to optimize inference accuracy are mainly divided into scaling up model size or quality, exploring the internal thinking mode of the large model, and developing multi-model collaboration strategies. Multi-model collaboration enhances inference accuracy by integrating multiple existing models, with significantly lower costs than the other two approaches. Consequently, this strategy has received widespread attention in recent years.

Existing multi-model collaborative reasoning methods are divided into full-response-level collaboration and token-level collaboration. Token-level collaboration has significant advantages over full-response-level collaboration in terms of token overhead and time cost. However, existing token-level collaboration methods face challenges such as insufficient filtering of low-confidence token noise and equalization of model contributions during the aggregation process. Therefore, designing an efficient token-level collaboration method to improve inference accuracy remains a key research challenge. Meanwhile, given the rapid development of the open-source large model ecosystem, private model deployment has gradually become a trend. However, locally deployed large models face performance bottlenecks and creating a need for accuracy improvement. Inspired by multi-model collaboration and the Internet's 'connectivity as a service' paradigm, interconnected models can cooperatively enhance reasoning accuracy. Thus, developing efficient token-level collaboration methods for systems with numerous interconnected models has important theoretical and practical value.

To solve the above problems, this paper presents a token-level multi-model parallel collaboration reasoning framework, namely DuetNet. In DuetNet, interconnected large language models collaboratively accomplish tasks through token-level coordination. The architecture improves inference accuracy through incremental aggregation of multi-model consensus. Specifically, during each inference step, DuetNet first employs a joint truncation strategy to minimize the introduction of low-confidence noise. Subsequently, during the aggregation phase, it calculates the cumulative logit scores of each candidate token by aggregating the logit value vectors, thereby mitigating confidence loss. Finally, the next token is selected using a Top-T stochastic sampling algorithm.

Experimental results show that DuetNet effectively improves the inference accuracy and reduces the user's uncertainty about model performance. Specifically, tests conducted on four inference datasets demonstrate that the model parallel collaborative reasoning under the DuetNet framework outperforms existing methods in terms of accuracy while incurring lower aggregation overhead. Within the DuetNet framework, the average inference accuracy of multi-model parallel collaboration exceeds that of single models by more than 21.44%, with the generation latency for individual Tokens increasing by only approximately 2 ms.