# 模型互联网中基于自我效能的Token级多模型协作

王建辉[1]，李哲涛[1]，石伟凡[1]，王泽平[1]，郑智润[2]，李成新[3]
（1.暨南大学信息科学技术学院，广东 广州 510632；2.亚洲大学人工智能系，韩国 水原 16499；
3.湘潭大学数学与计算科学学院，湖南 湘潭 411105）

**摘　要：**针对模型互联网中Token级协作在推理性能与开销难以兼顾的问题，提出一种基于自我效能的Token级多模型协作方法ConfiPara。首先，为解决现有Token级协作方法的高开销问题，设计一种具有退出机制的Token级多模型协作方法。其次，提出一种融合基模型自信度与信心可靠度的自我效能评估算法，用以判定退出时机；通过自我效能引导基模型在适当时转为独立推理，从而跳过冗余协作，在保证准确率的同时减少Token开销。实验结果表明，ConfiPara方法能以较小的准确率损失，显著降低Token消耗与推理时延。在单协作模型场景下，该方法仅以2.5%的准确率损失就能降低约21%的Token开销和最高75%的单Token生成时延。

**关键词：**大模型；模型互联网；Token级模型协作；退出机制；自我效能
**中图分类号：**TP181
**文献标志码：**A
**DOI：**10.11959/j.issn.1000–436x.2026033

# Token-level multi-model collaboration based on self-efficacy in AI-model network

Wang Jianhui[1], Li Zhetao[1], Shi Weifan[1], Wang Zeping[1], Zheng Zhirun[2], Li Chengxin[3]
1. College of Information Science and Technology, Jinan University, Guangzhou 510632, China
2. Department of Artificial Intelligence, Ajou University, Suwon 16499, Republic of Korea
3. School of Mathematics and Computational Science, Xiangtan University, Xiangtan 411105, China

**Abstract:** To address the trade-off between inference performance and cost in Token-level collaboration within the AI-model network, a self-efficacy-based Token-level multi-model collaboration method named ConfiPara was proposed. Firstly, a Token-level collaborative method with an exit mechanism was designed to mitigate the high overhead of existing approaches. Secondly, a self-efficacy assessment algorithm integrating the base model's confidence and reliability was introduced to determine the optimal exit timing. By leveraging self-efficacy to guide the base model in switching to independent inference at appropriate moments, redundant collaboration was skipped, thereby maintaining accuracy while reducing Token overhead. Experimental results demonstrate that the proposed ConfiPara method achieves a substantial reduction in Token consumption and inference latency with only a minor accuracy loss. In a single collaborative model scenario, the method reduces Token cost by approximately 21% and cuts per-Token generation latency by up to 75%, at the cost of only a 2.5% drop in accuracy.

**Keywords:** large model, AI-model network, Token-level model collaboration, exit mechanism, self-efficacy

**收稿日期：**2025–11–12；**修回日期：**2026–01–28
**通信作者：**李哲涛，liztchina@hotmail.com
**基金项目：**国家自然科学基金资助项目（No.W2411053, No.U23B2027）
**Foundation Items：**The National Natural Science Foundation of China (No.W2411053, No.U23B2027)

## 0 引言

近年来，大语言模型虽在多任务场景取得突破，但在复杂推理与知识泛化上仍存在固有局限。研究表明，即便模型规模持续扩大，单一模型在应对复杂推理、符号运算或深层常识理解等任务时仍面临性能瓶颈，这源于其知识边界、幻觉倾向及单一路径下自我纠错能力的不足[1-3]。为突破上述限制，研究范式正逐步从孤立模型优化转向多模型协同生态系统的构建，例如由异构模型组成的模型互联网系统[1, 4]，在该系统中，发起任务的主体被称为基模型，参与协同的其他模型则被定义为协作模型，二者的协同配合能够整合不同模型的差异化能力，显著提升复杂场景的问题解决能力。此类协作借鉴了群体智慧的思想，通过融合不同模型的优势与视角，以提升整体的推理准确性与鲁棒性[5]。

现有多模型协作研究大多聚焦输出级层面，如投票、讨论与辩论[6-8]，这类输出级层面的协作方法虽能有效提高推理准确率，但依赖多轮交互机制，易导致较高时延与开销。与之相比，Token级多模型协作方法能够实现更细粒度的协作，其核心在于每个Token生成步骤均进行多模型集体决策，不需要等待完整序列即可动态融合各模型实时输出[9]。此类方法避免了基于完整回复的多轮整合，在提高推理准确率的同时显著提升了推理效率，因而受到广泛关注。然而，现有Token级协作方法多采用全程协作策略，即在每个Token的生成过程中，所有模型均参与决策。由于在协作过程中各模型独立生成候选Token，因此总Token开销将随模型数量线性增长，导致通信开销相应增加。同时，每个Token的生成需要等待所有协作模型的响应，使推理时延显著增加。上述推理准确率与资源消耗之间的矛盾严重制约了Token级协作方法在模型互联网中的实际应用。因此，如何在尽可能维持推理准确率的前提下降低资源消耗，已成为推动该类方法落地的关键挑战。

为应对上述挑战，受心理学中自我效能理论[10]启发，本文提出了一种以模型信心为量化表征的自我效能驱动的Token级多模型协作方法ConfiPara（confidence-based parallel collaboration）。自我效能是指个体对自己是否有能力完成某一行为所进行的推测与判断，其高低直接影响个体是否选择独立应对挑战。自我效能理论强调个体在任务执行中基于信心水平自主调整行为方式：低自我效能的个体更可能寻求群体支持，高自我效能的个体则倾向于独立处理问题。基于上述理论机制，本文将模型的自我效能操作化为其推理信心度，并以此指导并联协作的触发与退出。具体而言，与自我效能理论所描述的行为模式类似，ConfiPara方法中设计了一种退出机制，即基模型在高自我效能时退出协作进行独立推理，反之则进行协作推理。这一机制模拟了个体从依赖协作到自主决策的行为转变过程，实现了基模型对协作模型依赖程度的动态调节。同时，退出机制的引入能从两方面提升系统效率：一是基模型在退出协作后独立推理，直接降低了总体Token消耗；二是避免了多模型间的通信与等待开销，从而有效降低了推理时延。此外，在ConfiPara方法中，基于自我效能来决策基模型退出协作的时机，从而在维持准确率的同时优化推理效率。

本文的主要贡献如下。

1)针对模型互联网中Token级协作的高开销问题，本文提出了一种基于自我效能的Token级协作方法ConfiPara。该方法通过引导基模型在高自我效能时进行独立推理，从而在可接受的准确率损失范围内，实现了Token开销的实质性降低。

2)在ConfiPara方法中，设计了一种融合自信度与信心可靠度的基模型自我效能评估算法。该算法综合考虑基模型在协作推理过程中的自信程度及其信心的可靠性，对基模型的自我效能进行量化建模，使其能够基于该效能指标自适应地判定是否进入独立推理阶段。

3) 实验结果表明，ConfiPara方法能在较小的准确率损失下显著降低Token开销和推理时延，展现出良好的成本效益。具体而言，在单协作模型场景下，该方法仅以2.5%的准确率损失就能降低约21%的Token开销。当基模型与协作模型性能相当时，ConfiPara展现出更显著的优势，其推理准确率提升约1%，同时Token开销相对降低约23%。此外，与采用全程协作策略的方法相比，ConfiPara方法下的单Token生成时延最多降低75%。

## 1 相关工作

### 1.1 多模型协作

在模型互联网中，现有多模型协作方法通常采用并联或层级式架构，在共享输入或上下文的条件

下，由多个模型并行或按顺序生成候选结果，并通过投票、加权或聚合机制形成最终输出。根据协作粒度的不同，相关工作可进一步划分为输出级协作和Token级协作。前者以模型的完整输出为交互单元，后者则将交互单元细化至下一个Token的概率分布。

现有输出级协作方法主要包括多数投票[6]、讨论[7]和辩论[8]等形式。其中，多数投票通过汇聚多个模型的预测结果并采用多数一致原则生成最终输出[6]；讨论机制依赖多轮模型交互与观点共享，常与投票策略结合使用，如CMD框架[7]；辩论方法则通过引入模型间的对立角色与交互质询来提升推理准确性[8]。尽管输出级协作在提升推理性能方面具有一定效果，但其通常依赖多轮模型交互，带来较高的Token与时延开销，因而更适用于对时延与成本约束相对宽松的应用场景。

不同于输出级协作需等待模型完整的输出，Token级协作是一类逐Token协作生成回复的方法。由于不同大模型之间的词汇差异，许多方法专注于对齐不同模型的分词器以优化集成效果[11-12]。Xu等[11]提出了基于词汇对齐的协作方法EVA，EVA通过学习词汇映射，将不同大模型的输出分布投影至统一空间，以实现细粒度集成。类似地，Huang等[12]基于相对表示理论，利用锚点与不同词汇的相对表示间接实现了词汇映射。上述两种方法[11-12]均依赖额外的训练过程，从而引入了相应的计算开销。对此，Yu等[13]提出了一种轻量化协作方法，其仅需一次矩阵乘法和一次分词操作即可实现Token级协作。在此基础上，Yao等[14]进一步从Token选择角度提升协作效率，提出了UniTe方法，该方法仅对各模型输出的Top-K Token进行协作，从而实现高效的Token级协作。针对聚合过程中可能引入的噪声问题，Wang等[15]提出了一种联合Top-K和Top-P的聚合策略以减少聚合过程中的噪声，从而优化推理准确率。总体来看，当前Token级协作方法仍以全程协作为主，其在提升推理性能的同时，也引入了较大的Token开销和推理时延。

### 1.2　模型自我效能评估

在模型决策过程中，自我效能通常被视为模型对其预测结果可靠性的主观评估，在相关研究中亦常以模型信心或置信度加以表述。现有工作多通过不确定性估计对模型自我效能进行建模，例如基于预测Token的概率、熵值等指标，对模型当前预测结果的可靠程度进行量化衡量[16]。尽管上述方法在一定程度上刻画了模型的决策信心，但其置信度校准问题仍被普遍认为是制约大模型实现可信推理的重要挑战。这是由于大模型存在一个显著风险：即使生成内容存在虚构或错误，仍可能表现出较高的置信水平。这一现象凸显了模型信心评估的内在复杂性，也推动了该领域广泛开展研究，旨在建立更可靠、可解释的置信度度量机制[16-17]。

针对大模型信心评估问题，已有研究探索了不同的校准机制[18-22]。Zhang等[18]提出了一种通过获取保真度来校准大模型信心的方法，通过将信心分为对问题的不确定性和对答案的保真度实现信心估计。Li等[19]提出了基于响应一致性构建的一致性图校准模型信心。Zhao等[20]通过设计事实抽取、反思和生成这3个步骤提示实现模型在不确定时主动表达“我不确定”。Zhang等[21]提出了一种基于预训练数据分布差异估计置信度的方法，能够在无标签情况下对模型进行校准。上述方法侧重于估计或引出个体信心，没有充分利用集体智慧。对此，Yang等[22]引入多模型协商机制，利用多模型的集体智慧提升信心校准性能，通过小组讨论、评分和多数投票实现协作信心校准。

总体而言，现有关于模型信心的研究主要聚焦于单模型信心评估，对Token级协作场景下的信心建模关注相对有限。此外，当前Token级多模型协作方法普遍采用全程协作策略。在该策略下，无论各模型对当前任务的实际信心水平如何，均需全程参与每一生成步骤，导致计算资源消耗与推理时延随协作规模增加而显著上升。针对上述问题，本文设计了一种基于自我效能的Token级协作方法，通过模型自我效能判断何时退出协作进行独立推理，从而在保证推理准确率的同时优化推理效率。

## 2　ConfiPara方法

### 2.1　系统模型

模型互联网中多模型协作架构如图1所示，主要由模型互联平台、模型实体和用户3个部分构成。模型互联平台作为系统的核心组成部分，承担着模型的动态接入与退出、统一管理以及任务调度等关键职能，是连接各模型并实现高效协作的媒介。在系统运行过程中，模型作为基本功能实体，通过模型互联平台被有效组织与调度，从而为用户

提供集成化的服务支持。用户则作为任务的发起方，将具体任务提交至模型互联平台，由平台协调底层模型资源完成处理与响应。

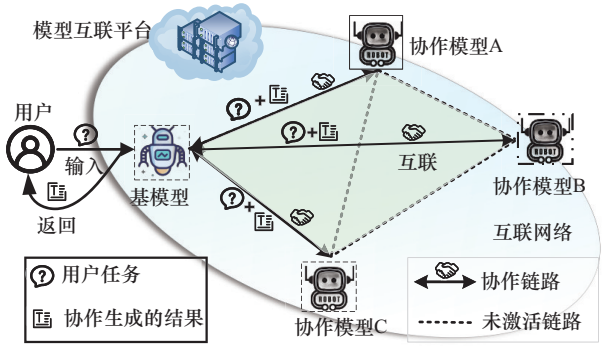



图1　模型互联网中多模型协作架构

在上述架构中，模型实体根据其角色划分为基模型与协作模型。用户输入的任务首先被提交至一个基模型，该模型随后将依据预设的协作规则，与一个或多个协作模型通过互联网进行交互，通过聚合多个模型的响应共同完成该任务。上述过程的具体步骤如下：首先，基模型收到用户提交的任务并将该任务转发给相应的协作模型；其次，基模型与协作模型通过预设的协作方式共同完成该任务；最后，基模型收到任务的最终结果并转发给相应的用户。上述协作过程中的聚合操作在基模型侧完成，这意味着协作模型需将中间结果发送至基模型，在基模型侧进行聚合并决策后续步骤。

### 2.2　具有退出机制的 Token 级多模型协作

为应对模型互联网中现有 Token 级协作方法的高 Token 开销问题，本文在 ConfiPara 方法中设计了一种受自我效能理论启发、具有退出机制的 Token 级多模型协作推理算法，整体流程如图 2 所示。该算法将推理过程划分为两个阶段，即多模型协作推理阶段和基模型独立推理阶段。具体而言，算法将基模型的自我效能初始化为 0，当基模型接收到用户任务后，进入多模型协作推理阶段，通过观察多模型协作过程中的数据更新自我效能；当基模型的自我效能超过阈值且任务尚未结束时，将进入基模型独立推理阶段，基模型基于多模型协作推理阶段生成的内容，独立完成后续的推理。

由上述过程可见，ConfiPara 方法允许基模型在满足条件时退出协作并转为独立推理。该设计实现了三重优化：参与模型数量的动态减少直接降低了 Token 开销，从而使通信开销相应降低；退出协作后不需要等待多方响应，可实现推理时延降低；以自我效能为决策依据确保退出时机合理，在维持准确率的同时提升效率。

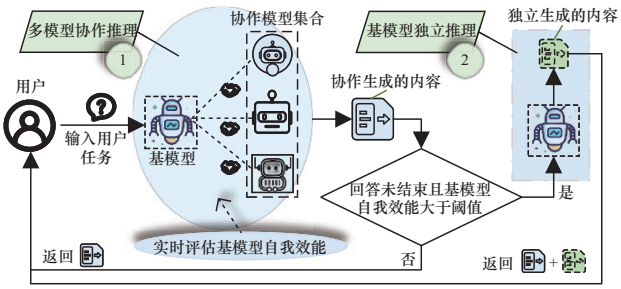



图2　ConfiPara 方法中具有退出机制的 Token 级多模型协作方法整体流程

令 $\mathcal{M}=\{B_0,C_1,C_2,\cdots,C_N\}$ 表示基模型和 $N$ 个协作模型的集合，其中，$B_0$ 为基模型，$C_n$ 表示第 $n$ 个协作模型。令 $Q$ 为任务集合，$\tau\in Q$ 表示用户提交的任务，$A_{\text{true}}$ 表示该任务的正确答案。令 $A_\tau$ 和 $E_\tau$ 表示多模型协作推理生成的推理答案和 Token 开销，$\delta(x,y)$ 为克罗内克函数，在 $x=y$ 时取值为 1，否则为 0。以推理准确率与 Token 开销两个指标评价本文方法的性能，具体指标定义如式(1)所示。

$$\begin{cases}\max \dfrac{\sum\limits_{\tau\in Q}\delta\left(A_{\text{true}},A_\tau\right)}{|Q|}\\ \min \dfrac{\sum\limits_{\tau\in Q}E_\tau}{|Q|}\end{cases}\tag{1}$$

其中，第一项衡量平均推理准确率，第二项衡量平均 Token 开销。

具有退出机制的 Token 级多模型协作的具体过程如算法 1 所示。算法初始化协作模型集合为 $H=\{1,2,\cdots,N\}$，协作结果 $R$ 为空。在该算法中，基模型独立推理与 Token 级多模型协作推理的唯一区别在于聚合对象：前者仅聚合基模型自身的输出分布，后者聚合基模型与协作模型的输出分布。当协作模型集合 $H$ 为空时，即独立推理的特殊情形，故本节不再单独描述独立推理阶段。

**算法 1**　具有退出机制的 Token 级多模型协作算法

**输入**　$\mathcal{M}$、$\tau$、$\gamma$ 和 $\theta$

**输出**　协作结果 $R$

1)$H=\{1,2,\cdots,N\}$，$R=$""

2)while 生成未结束 or 未达到最大生成 Token 数量 do

3)　$L_0=B_0(\tau,\gamma)$

4)　$L_n=C_n(\tau,\gamma)$，$n\in H$

5)　$u=\mathrm{COLL}\left(\{L_0,L_1,\cdots,L_N\}\right)$

6)　$\tau=\tau+u$，$R=R+u$

7)　评估基模型自我效能 $T_0$

8)　如果 $T_0>$ 阈值 $\theta$，那么 $H=\varnothing$

9)end while

10)return $R$

与基模型独立推理类似，Token 级多模型协作推理同样遵循自回归生成过程。在该协作推理阶段的每一个 Token 生成步骤中，各模型依据协作生成的公共 Token 序列进行下一步预测，具体流程如下。在未达到终止条件前，各模型首先基于当前公共 Token 序列并行地进行自回归推理，生成下一个 Token 的概率分布（第3~4行）。随后，对各模型输出的概率分布进行聚合并采样，以确定下一个公共 Token，并将其追加至公共 Token 序列中（第5~6行）。上述生成、聚合和采样过程循环进行，直至生成结束或达到最大生成 Token 数量限制。同时，基模型在协作生成过程中对自我效能 $T_0$ 进行动态评估，当效能大于阈值 $\theta$ 时，基模型将转入独立推理阶段（第7~8行）。由于协作过程以解码后的字符串作为输入，因此该方法具备与不同模型分词器的兼容性，从而能够集成异构模型进行协作。

上述多模型进行 Token 级协作生成下一个公共 Token $u$ 的过程，可形式化表述如下。

首先，基模型生成下一个 Token 的候选分布 $L_0$ 为

$$L_0=B_0(\tau,\gamma) \tag{2}$$

其中，$B_0(\tau,\gamma)$ 表示基模型根据输入任务 $\tau$ 和模型控制参数 $\gamma$（如 Top-P 和 Top-K 的设置）输出候选分布。

然后，每个协作模型生成下一个 Token 的候选分布。其中，协作模型 $L_n$ 生成的候选分布可表示为

$$L_n=C_n(\tau,\gamma),n\in H \tag{3}$$

其中，$C_n(\tau,\gamma)$ 表示协作模型 $n$ 基于输入 $\tau$ 和控制参数 $\gamma$ 输出的候选分布。

最后，聚合基模型和协作模型的候选分布并选择出所有模型共用的下一个 Token $u$，该过程可表示为

$$u=\mathrm{COLL}\left(\{L_0,L_1,\cdots,L_N\}\right) \tag{4}$$

其中，$\mathrm{COLL}(\cdot)$ 表示采样函数，例如选择累积概率最高的候选 Token[14-15]。

Token 级多模型协作推理示例如图 3 所示。在每一生成步骤中，前序 Token 作为共享上下文并行输入各模型，各模型独立生成候选 Token，随后通过聚合与采样确定公共 Token，并将其附加至上下文以驱动下一步自回归生成。基模型自我效能作为协作触发与退出的依据贯穿上述过程，但为突出 Token 级协作流程，其内部计算未在图 3 中显式展示。

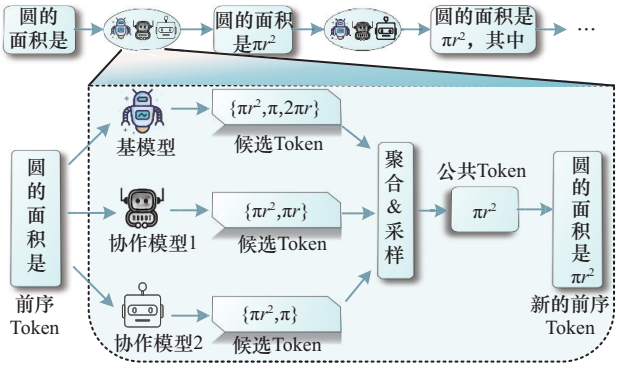



图3　Token级多模型协作推理示例

## 2.3　基模型自我效能评估

在 ConfiPara 方法中，基模型的自我效能通过模型信心进行量化，二者呈正相关性。然而，类似于认知系统中的评估偏差，基模型在推理中也可能出现信心失准。过度自信会使其在应协作时转入独立推理，增加任务失败风险；而信心不足则会推迟退出协作，导致 Token 开销上升。

针对上述挑战，本文提出了一种融合基模型自信度和信心可靠度的基模型自我效能评估算法，如图 4 所示。该算法的核心思想是融合基模型的全局自信度与信心可靠度，以更准确地评估其真实信心水平，从而抑制因模型过度自信和信心不足而产生的误判。

鉴于仅凭少数几次交互往往难以准确评估基模型的自我效能，本文引入了图 4 中的观察窗口策略。该策略要求至少累积一个最小观察窗口量

（$\omega$）的交互数据，方可更新基模型自信度。为便于后续阐述评估方法细节，给出相关定义如下。

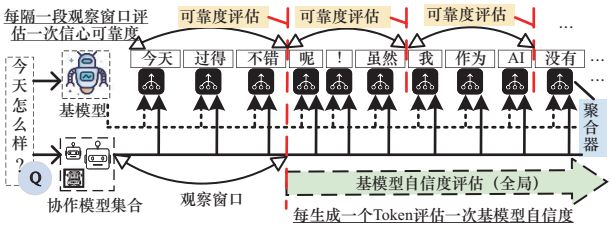



图4　基模型自信度和信心可靠度评估示意

**定义 1**　Token的优先级。令$\Theta_m$表示模型$m$在当前步骤生成的按概率值从大到小排序的候选Token集合，Token $u \in \Theta_m$在模型$m$当前推理步骤的优先级即指$u$在$\Theta_m$中的下标，若$u \notin \Theta_m$，则下标为无穷大。本文定义Token $u$在模型$m$的当前生成步骤中对应的下标越小，则优先级越高，即该模型在下一步选择此Token的概率越高。

**定义 2**　置信度优势。令$\hat{u}$表示一个推理步骤中通过Token级多模型协作推理生成的Token，若$\hat{u}$在基模型中的优先级高于协作模型$C_n$，则表示基模型在当前推理步骤的置信度高于$C_n$，否则低于$C_n$。

### 2.3.1　基模型自信度评估

令$D_t = \{\pi_1, \pi_2, \cdots, \pi_t\}$表示第$t$个生成步骤以来采集的信心证据。其中，$\pi_t$表示在第$t$个生成步骤中采集的信心证据，其取值为该步骤中置信度低于基模型的协作模型数量。依据定义2，公共Token在基模型中的优先级直接反映了其与群体共识的一致性。该优先级越高表示基模型与群体共识越一致，因此基模型的信心增加，反之则信心降低。鉴于此，本文采用集合$D_t$作为衡量自信度的证据。

若将单次公共Token的生成过程视为一次伯努利试验，基模型优势则表示“成功”，否则为“失败”。那么，多模型协作逐Token生成回复的过程则可视为一系列独立的伯努利试验。在该设定下，成功次数越多，表明基模型的置信水平越高，其越可能在后续的独立推理中正确完成任务。然而，真实的成功概率$P$本身具有不确定性。为此，本文引入Beta分布建模该成功概率的信念，即$P$~Be$(s,f)$，$s$和$f$分别表示成功系数和失败系数。Beta分布能够灵活表达对成功概率的先验认知，其可通过新的实验结果进行贝叶斯更新，并使用该分布的期望值估计成功概率。

为适应多模型协作场景，本文对“成功”与“失败”的判定规则进行了扩展。具体地，在每一公共Token生成步骤中，若基模型的置信度高于某一协作模型，则计为一次“成功”；反之则计为一次“失败”。在此基础上，单个生成步骤中的“成功”总数定义为基模型置信度超过的协作模型数量（即$\pi_i$），“失败”总数则为置信度不高于基模型的协作模型数量（即$N-\pi_i$）。每生成一个公共Token，成功系数$s$随之增加$\pi_i$，失败系数$f$增加$\rho(N-\pi_i)$。在第$t$个Token生成步骤时，累计的$s$和$f$分别为

$$s = \sum_{i=1}^{t} \pi_i \tag{5}$$

$$f = \rho \sum_{i=1}^{t} (N - \pi_i) \tag{6}$$

其中，$\rho > 1$为惩罚因子，初始时，$s$和$f$均设为1。

最后，基模型的自信度可通过计算上述Beta分布的数学期望来估计，即式(7)。该期望值反映了基模型在协作中的平均成功概率，可作为其自信水平的度量。

$$\mathrm{ST}_0 = \boldsymbol{E}(P) = \frac{s}{s+f} \tag{7}$$

### 2.3.2　基模型信心可靠度评估

考虑到模型在单一Token上的优异表现可能具有偶然性，无法真实反映其持续的推理能力，本文引入了信心可靠度的概念。由图4可以看出，细粒度的基模型自信度评估在每一Token生成后即时进行，其证据来源于模型在当前Token上的瞬时表现；而粗粒度的基模型信心可靠度则将评估尺度扩展至一个包含多个Token的观察窗口，以模型在此窗口内的整体表现作为证据。通过将信心证据从单个Token扩大到一个观察窗口，可以对基模型在一段时间内的表现进行平滑评估，有效过滤随机波动，从而判断其信心水平是否稳定可靠。在该示例中设定观察窗口大小为3，因此，在生成第3个Token后，将基模型在该观察窗口内的整体表现作为证据评估基模型的信心可靠度。由于此时基模型的自我效能低于阈值，观察窗口进行无重叠滑动。随后，在完成后续3个Token的生成后，再次对基模型的信心可靠度进行评估。

在多模型协作推理过程中，单一模型在不确定情形下往往难以准确评估其自身判断的可靠性。为收集衡量基模型信心可靠度的证据，本文采用了源

于社会认同理论的启发式原则。该理论指出个体在缺乏充分信息时，通常通过与群体判断的一致性来进行自我校准：当个体决策与群体共识保持一致时，其主观信心倾向于增强；反之，则会降低对自身判断的可靠度。形式化地，该过程可表示为

$$J_n = J_o + f\left(\mathrm{sim}\left(G, I\right)\right) \tag{8}$$

其中，$G$表示群体共识，$I$为个体判断，$J_o$和$J_n$分别表示个体在更新前后的主观信心水平，$\mathrm{sim}(\cdot,\cdot)$表示一致性度量函数，$f(\cdot)$是一个单调非减函数。

在模型互联网中的Token级多模型协作推理场景中，上述启发式原则可进一步映射为对模型信心可靠度的评估。具体而言，基模型的局部信心度对应个体判断$I$，而协作模型的平均信心水平则刻画群体共识$G$。若基模型的局部信心度与协作模型群体的平均信心水平相当或更高，可认为其判断与群体共识一致，从而具有较高的信心可靠度，此时可靠性系数$s_w$递增；反之，则表明其判断可能源于过度自信，此时不可靠系数$f_w$递增。基于以上论述，评估基模型的信心可靠度的具体过程如下。

首先，计算每个模型在当前观察窗口的自信度。令$e_s^m$和$e_f^m$分别表示模型$m$在该窗口的成功系数和失败系数，它们在每个新的窗口都初始化为1。基于式(5)和式(6)更新该窗口的成功系数$e_s^m$和失败系数$e_f^m$。由此，可通过式(7)得到模型$m$在当前窗口的局部自信度$W_m$，即$W_m = \dfrac{e_s^m}{e_s^m + e_f^m}$。

其次，通过式(9)计算所有协作模型在当前窗口的平均信心度$\bar{W}$为

$$\bar{W} = \frac{\sum_{n=1}^{N} W_n}{N} \tag{9}$$

然后，通过式(10)计算当前窗口基模型的信心是否可靠。

$$\begin{cases} \text{可靠}, & W_0 - \bar{W} > -\epsilon \\ \text{不可靠}, & \text{其他} \end{cases} \tag{10}$$

其中，$W_0$表示基模型在当前窗口的局部自信度，较小的常数$\epsilon > 0$表示相当程度因子。由式(10)可见，$W_0 - \bar{W} > -\epsilon$表示基模型的局部信心度与协作模型群体的平均信心水平相当或更高，此时认为基模型的信心可靠。

本文将信心可靠度也建模为一个Beta分布。与式(5)、式(6)相似，可靠性系数$s_w$和不可靠系数$f_w$的更新规则可表示为式(11)。当前窗口基模型的信心可靠，则$s_w$自增1；否则$f_w$自增$\rho$。

$$\begin{cases} s_w = s_w + 1, & \text{可靠} \\ f_w = f_w + \rho, & \text{其他} \end{cases} \tag{11}$$

其中，参数$\rho$即式(6)所定义的惩罚因子，二者取值相同；初始化时，$s_w$和$s_f$都赋值为1。

最后，类比于式(7)，基模型的信心可靠度可通过式(12)得到，即

$$\mathrm{RT}_0 = \frac{s_w}{s_w + f_w} \tag{12}$$

### 2.3.3　基模型自我效能评估

首先，评估基模型自信度和信心可靠度，然后综合二者得到基模型的自我效能。本文通过基模型的自信心与信心可靠度的乘积来量化自我效能$T_0$，即式(13)。该设计通过乘法耦合使任一因子偏低均会抑制最终估值，从而有效防止模型因过度自信而过早退出协作进行独立推理。

$$T_0 = \mathrm{ST}_0 \times \mathrm{RT}_0 \tag{13}$$

初始化时，令基模型自信度评估的最小观察窗口参数$\omega_1 = \omega$，基模型信心可靠性评估的观察窗口大小参数$\omega_2 = \omega$。每生成一个公共Token后，$\omega_1$和$\omega_2$都自减1。

对于基模型自信度的评估，若当前协作生成的公共Token数量小于最小观察窗口$\omega_1$，则仅收集信心证据；若$\omega_1$不为正，则表明自信度评估所收集的证据量不低于最小观察窗口，此后在每个生成步骤都将更新基模型自信度。

对于基模型的信心可靠度评估，若$\omega_2$为0，则表明可靠性评估所收集的证据量恰好等于观察窗口大小，此时将更新基模型的信心可靠度；若$\omega_2 > 0$，则仅收集用于评估可靠度的证据数据。

基模型的自我效能$T_0$在每个生成步骤中基于式(13)进行更新。若$T_0$不超过阈值$\theta$，那么观察窗口将进行不重叠平移，即$\omega_2$重新赋值为$\omega$，这意味着还需进行信心可靠度评估。若综合信心$T_0$超过阈值$\theta$，那么认为基模型具有较高的信心，可以独立推理正确，此时基模型退出协作进行独立推理。基模型自我效能评估算法如算法2所示。

**算法2**　基模型自我效能评估算法

**输入**　最小观察窗口$\omega$和阈值$\theta$

**输出** 基模型自我效能 $T_0$

1)$s_w = f_w = s = f = 1$

2)$\omega_1 = \omega_2 = \omega$ // $\omega_1$和$\omega_2$分别表示自信度评估最小观察窗口和可靠性评估观察窗口

3) while 生成未结束 or 未达到最大生成Token数量 do

4) Token级协作生成当前步骤的下一个Token

5) 基于式(5)和式(6)更新$s$和$f$

6) $\omega_1 = \omega_1 - 1$，$\omega_2 = \omega_2 - 1$

7) 如果$\omega_1 \leqslant 0$，更新基模型自信度$\mathrm{ST}_0$

8) if $\omega_2 == 0$ then

9) 计算每个模型的局部窗口自信度

10) 基于式(11)更新$s_w$和$f_w$

11) 更新基模型信心可靠度$\mathrm{RT}_0$

12) end if

13) 更新基模型自我效能 $T_0$

14) 如果$T_0 \leqslant \theta$，那么$\omega_2 = \omega$；否则基模型开始独立推理

15) end while

## 3 实验结果分析

### 3.1 实验设置

实验选定的4款开源大模型如表1所示，实验环境为配备4×80 GB A100的服务器，各模型分别独立部署于单张GPU，以支持Token级协作推理的并行执行。本文实验选定的测试数据集及对应的提示词模板如表2所示，从每个数据集中随机抽取50个测试样本，在所有实验中统一采用相同的提示词设置，以确保评测结果的公平性与可复现性。

本文的对比方法如下。

基模型：仅基模型独立推理，并且其Token生成过程与ConfiPara中基模型独立推理一致。

**表1 实验选定的4款开源大模型**

| 简称 | 模型 | 发布者 |
|---|---|---|
| DS | deepseek-llm-7b-chat | 深度求索 |
| Q2 | Qwen2.5-14B-Instruct | 阿里巴巴集团 |
| H3 | Hunyuan-7B-Instruct | 腾讯 |
| G4 | glm-4-9B-chat | 智谱AI |

UniTe[14]：一种基于Top-K截断的Token级全程协作方法，参照原工作设置，设定$K$值为10。

DuetNet[15]：一种采用联合Top-K和Top-P截断策略的Token级全程协作方法，参照原工作设置，设定$K$值为10，$P$值为0.75。

UPara-N：一种基于固定协作Token数量的Token级协作方法，该方法延续UniTe的协作范式，但限定仅在生成前$N$个Token时启动多模型协作，后续生成阶段则由基模型独立完成。

DPara-N：与UPara-N类似，但是该方法的协作范式与DuetNet一致。

为全面评估ConfiPara方法的兼容性，本文基于上述两种全程协作范式分别构建了方法实例：ConfiPara-U（协作范式基于UniTe）与ConfiPara-D（协作范式基于DuetNet）。为系统评估ConfiPara的性能表现并量化其优势，本文基于式(1)定义的两个评价指标，采用推理准确率损失与Token开销降低比作为ConfiPara方法的核心评价指标。在性能比较过程中，由于不同方法采用的协作范式存在本质差异，本文严格遵循同类范式对比原则。例如，ConfiPara-D仅与基于相同协作范式的DuetNet及DPara-N进行比较。

推理准确率损失。本文定义推理准确率损失为全程协作方法的推理准确率与当前方法准确率之差，用以衡量因提前退出协作所带来的潜在性能代价，即

**表2 实验选定的测试数据集及对应的提示词模板**

| 简称 | 数据集 | 数据类型 | 提示词模板 |
|---|---|---|---|
| D1 | TrustfulQA[23] | 选择题 | 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. |
| D2 | MMLU[24] | 选择题 | 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. |
| D3 | GSM8K[25] | 数学推理题 | Please answer the following question: {question} You must reiterate your answer numerically at the end of the response. |
| D4 | BoolQ[26] | 判断题 | 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. |

$$\text{loss} = \text{Acc}_{\text{full}} - \text{Acc}_{\text{exit}} \tag{14}$$

其中，$\text{Acc}_{\text{full}}$表示全程协作方法的推理准确率，即DuetNet或UniTe；$\text{Acc}_{\text{exit}}$表示具备退出机制方法的推理准确率。

Token开销降低比。本文定义Token开销降低比为全程协作方法的Token开销与对比方法Token开销的相对减少比例，即

$$\text{save} = \frac{\text{Cost}_{\text{full}} - \text{Cost}_{\text{exit}}}{\text{Cost}_{\text{full}}} \tag{15}$$

其中，$\text{Cost}_{\text{exit}}$和$\text{Cost}_{\text{full}}$分别对应具备退出机制方法与全程协作方法的单问题平均Token开销。

### 3.2　参数影响分析

本节分析观察窗口大小$\omega$和效能阈值$\theta$对ConfiPara性能的影响，单模型推理准确率如表3所示。由表3可以看出，模型DS的表现最差，而Q2的表现最优。

**表3　　单模型推理准确率**

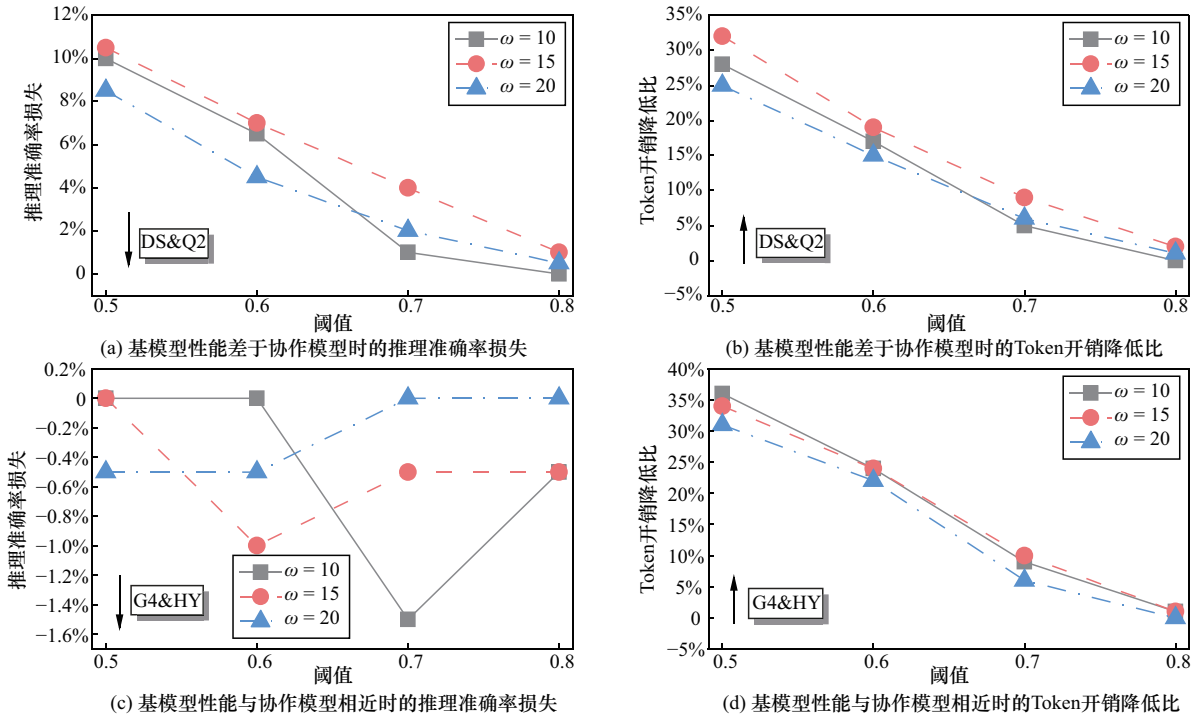
| 模型 | D1 | D2 | D3 | D4 | 平均 | 排名 |
|---|---|---|---|---|---|---|
| Q2 | 86% | 84% | 94% | 82% | 86.5% | 1 |
| HY | 68% | 66% | 84% | 74% | 73.0% | 2 |
| G4 | 60% | 68% | 82% | 80% | 72.5% | 3 |
| DS | 50% | 46% | 60% | 70% | 56.5% | 4 |

基于表3所示结果，本文分析了两种典型场景下不同$\omega$和$\theta$参数对ConfiPara-U性能的影响，包括基模型性能明显低于协作模型（即DS&Q2）和基模型与协作模型性能相近（即G4&HY）场景。不同$\omega$和$\theta$下，ConfiPara-U与UniTe相比的性能表现如图5所示。

由图5可以看出，随着效能阈值的增加，两种场景的Token开销降低比都呈现近乎线性的减少趋势。因此，为均衡推理准确率和Token开销，效能阈值$\theta$不宜过大。但是，当$\theta = 0.5$时，DS&Q2的准确率损失大于8%，这说明阈值也不宜太小。综合考虑二者，本文设置$\theta = 0.6$。进一步分析$\omega$参数对性能的影响可以发现，当$\theta = 0.6$时，DS&Q2组合在$\omega$值较大时所损失的推理准确率更低，而G4&HY组合实现了性能的正向增长。该结果表明较大的$\omega$值整体上更具优势。因此，综合考虑推理准确率损失和Token开销降低比，在后续实验中，本文设置$\omega = 20$、$\theta = 0.6$。

### 3.3　单协作模型场景性能分析

本节对单个协作模型场景下不同方法的性能进行了分析，不同场景下不同方法的表现如表4所示，其中，“loss”表示推理准确率损失，“save”表示Token开销降低比，“组合”列遵守“基模型&




(a) 基模型性能差于协作模型时的推理准确率损失　(b) 基模型性能差于协作模型时的Token开销降低比

(c) 基模型性能与协作模型相近时的推理准确率损失　(d) 基模型性能与协作模型相近时的Token开销降低比

图5　不同$\omega$和$\theta$下，ConfiPara-U与UniTe相比的性能表现

表 4 不同场景下不同方法的表现

| 场景 | 组合 | 基模型 | DuetNet | | DPara-20 | | ConfiPara-D | | UniTe | | UPara-20 | | ConfiPara-U | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 准确率 | 准确率 | 开销 | loss↓ | save↑ | loss↓ | save↑ | 准确率 | 开销 | loss↓ | save↑ | loss↓ | save↑ |
| 基模型性能较差 | DS&Q2 | 56.5% | 78.5% | 237.1 | 13% | 44% | 5% | 15% | 80.5% | 233.4 | 16% | 45% | 5% | 15% |
| | DS&HY | 56.5% | 63.5% | 199.8 | 4% | 43% | 3% | 37% | 69.5% | 273.5 | 9% | 50% | 6% | 12% |
| | G4&Q2 | 72.5% | 84.0% | 267.9 | 6% | 52% | 1% | 11% | 83.0% | 249.7 | 4% | 50% | 4% | 23% |
| | HY&Q2 | 73.0% | 87.5% | 283.8 | 8% | 35% | 4% | 19% | 84.0% | 290.6 | 7% | 34% | 3% | 22% |
| 基模型与协作模型性能相近 | G4&HY | 72.5% | 78.5% | 266.9 | 5% | 54% | 0 | 15% | 76.5% | 240.5 | −1% | 50% | −1% | 22% |
| | HY&G4 | 73.0% | 78.5% | 266.9 | 0 | 37% | 2% | 31% | 76.5% | 240.5 | −1% | 31% | −1% | 24% |
| **平均** | | **67.3%** | **78.4%** | **253.7** | **6%** | **44%** | **2.5%** | **21%** | **78.3%** | **254.7** | **5%** | **43%** | **2.6%** | **20%** |
| 基模型性能更优 | Q2&DS | 86.5% | 78.5% | 237.1 | −6% | 38% | −4% | 25% | 80.5% | 233.4 | −4% | 39% | −3% | 39% |
| | HY&DS | 73.0% | 63.5% | 199.8 | −8% | 18% | −1% | 5% | 69.5% | 273.5 | −2% | 36% | −3% | 27% |

协作模型”的固定结构，例如DS&Q2表示基模型为DS，协作模型为Q2。

首先，当基模型性能较差时，两种全程协作方法（即DuetNet和Unite）都有效提高了基模型的推理准确率，最大提升了约24%，这表明Token级协作推理可有效提高基模型的推理准确率。同时，也可以发现ConfiPara在损失较小的推理准确率的同时可以有效降低Token开销。与DuetNet相比，ConfiPara-D的推理准确率降低了1%~5%，但是Token开销降低比达到11%~37%。与UniTe相比，ConfiPara-U的推理准确率降低了3%~6%，Token开销降低比达12%~23%。此外，与DPara-20和UPara-20相比，ConfiPara的推理准确率损失更小；尤其在DS&Q2组合下，DPara方法的推理准确率损失达到10%以上，明显高于ConfiPara方法。

其次，当基模型与协作模型性能相近时，全程协作的方法依旧有效提升了推理准确率，提升幅度约为5%。此外，可以发现ConfiPara方法下的准确率损失更少，同时Token开销也得到了有效的降低。当二者性能相近时，具有退出机制的方法甚至使推理准确率高于全程协作的方法。例如，G4&HY组合下，ConfiPara-U的推理准确率提高了1%，同时Token开销降低了约23%。导致该现象的一个潜在原因在于参与协作的基模型在性能上较为接近，在经过短期的共识积累后，若基模型退出协作并转为独立推理，其推理准确率可能与全程保持协作时差异不大。此外，独立推理过程避免了其他模型的干扰，模型的判断可能更为稳定，从而有助于提升其推理的正确性。

最后，本文也对基模型性能优于协作模型的场景开展了实验。当基模型性能更优时，协作并不能带来推理准确率的提升，具有退出机制的方法其推理准确率高于全程协作方法。上述现象表明，协作并非始终带来性能提升，其在引入共识增益的同时，也可能产生协作干扰。模型之间的共识有利于减少推理错误，但是模型之间的互相干扰则可能导致推理错误。基于本文的实验发现，为保证效益，较优的基模型不宜与较差的协作模型进行Token级协作。

综合上述3个典型场景的实验可知，在基模型性能不优于协作模型的场景下，ConfiPara方法表现出良好的有效性。与全程协作方法相比，ConfiPara方法平均推理准确率虽然降低约2.5%，但是Token开销相对降低了约21%。此外，ConfiPara的推理准确率损失明显低于采用固定协作Token数量规则的方法（即DPara和UPara），该结果也论证了本文ConfiPara方法的信心评估具有一定的效果。需要明确的是，本文遵循以推理准确率为首要优化目标的原则。由表4结果可见，尽管DPara与UPara方法在降低Token消耗方面表现出显著优势，但其伴随的推理准确率下降幅度过大，构成了难以接受的性能代价。此外，固定协作次数的方法存在“一刀切”的局限，无法适应不同问题的复杂性；而ConfiPara通过效能阈值实现动态退出。ConfiPara的这种“按需协作”机制避免了资源浪费与过早终止，更具稳健性与效率。

### 3.4 多协作模型场景性能分析

为进一步探索ConfiPara的性能边界，本节分

析了多个协作模型场景下不同方法的性能。与表4类似，DS&G4&Q2表示基模型DS，协作模型为G4和Q2，其余情形类推。

### 3.4.1 基模型性能较差场景

基模型较差时不同方法的表现如表5所示。基于表5，本节有以下发现。

1) 在推理准确率方面，具有Token级协作策略的方法明显优于基模型独立推理，最大提升约为26%。该结果进一步表明了Token级协作可有效提高基模型的推理准确性。

2) 采用全程协作的DuetNet与UniTe方法，其推理准确率随协作模型数量增加而提升，但同时也导致了计算开销的增长。然而，DS&G4&HY组合的准确率显著偏低，这表明协作效率的提升有赖于模型间的有效互补，而非简单的数量堆叠。

3) 在多协作模型场景中，ConfiPara的推理准确率损失达到7.5%，显著高于单协作模型场景的2.5%。出现该现象的可能原因是协作模型数量的增加使基模型更容易出现过度自信，进而导致其过早进行独立推理。因此，在多协作模型场景中，若用户看重推理准确率，则可以设置高效能阈值优化推理准确率。尽管如此，与采用固定协作Token数量的方法相比，ConfiPara也展示了明显的优势，平均提升推理准确率达10%。

### 3.4.2 基模型性能较优场景

基模型较优时不同方法的表现如表6所示。基于表6，本节有以下发现。

1)与DPara和UPara相比，ConfiPara的准确率损失相近，但Token开销降低比却更小。这说明在基模型较优时，提前退出并不会显著影响其独立推理的准确性。然而，在实际场景中，用户缺乏对模型优劣的全面认知，因此若选择的基模型性能较差，那么采用固定协作Token数量的方法将会导致推理准确率大幅下降。

2)在基模型性能较优时，全程协作方法（即DuetNet和UniTe）依旧可以有效提升基模型的推理准确率，最大提升幅度达到10%。该结果进一步论证了Token级多模型协作的有效性。

3) 与表5所示的基模型性能较差的结果相比，当基模型性能较优时，ConfiPara方法损失的推理准确率更少且更节省Token开销。与DuetNet和UniTe相比，ConfiPara方法平均损失的推理准确率低于4%，但是平均Token开销降低了至少32%。特别地，与UniTe相比，ConfiPara-U实现了损失2.1%左右的推理准确率，但Token开销降低了约36.5%。该结果表明在基模型性能较优时，ConfiPara可以实现在多协作模型场景中损失较少的推理准确率的同时大幅度降低Token开销。

**表5　基模型较差时不同方法的表现**

| 组合 | 基模型 | DuetNet | | DPara-20 | | ConfiPara-D | | UniTe | | UPara-20 | | ConfiPara-U | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 准确率 | 准确率 | 开销 | loss↓ | save↑ | loss↓ | save↑ | 准确率 | 开销 | loss↓ | save↑ | loss↓ | save↑ |
| DS&G4&Q2 | 56.5% | 82.5% | 324.0 | 18% | 55% | 7% | 21% | 80.0% | 305.0 | 13% | 53% | 5% | 15% |
| DS&HY&Q2 | 56.5% | 82.5% | 366.2 | 17% | 58% | 6% | 18% | 81.5% | 388.1 | 18% | 60% | 5% | 12% |
| DS&G4&HY | 56.5% | 68.5% | 285.4 | 5% | 52% | 5% | 34% | 78.0% | 334.2 | 11% | 57% | 4% | 18% |
| DS&G4&HY&Q2 | 56.5% | 82.0% | 470.9 | 19% | 64% | 12% | 23% | 83.0% | 442.9 | 17% | 62% | 6% | 18% |
| 平均 | **56.5%** | **78.8%** | **361.6** | **14.7%** | **57%** | **7.5%** | **24%** | **80.6%** | **367.5** | **14.7%** | **58%** | **5%** | **15.7%** |

**表6　基模型较优时不同方法的表现**

| 组合 | 基模型 | DuetNet | | DPara-20 | | ConfiPara-D | | UniTe | | UPara-20 | | ConfiPara-U | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 准确率 | 准确率 | 开销 | loss↓ | save↑ | loss↓ | save↑ | 准确率 | 开销 | loss↓ | save↑ | loss↓ | save↑ |
| G4&HY&Q2 | 72.5% | 83.5% | 396.0 | 4% | 63% | 3% | 23% | 81.0% | 380.3 | 4% | 64% | 1% | 30% |
| G4&DS&Q2 | 72.5% | 82.5% | 324.0 | 3% | 59% | 2% | 30% | 80.0% | 305.0 | 1% | 59% | 1% | 38% |
| G4&DS&HY&Q2 | 72.5% | 82.0% | 470.9 | 6% | 67% | 3% | 37% | 83.0% | 442.9 | 6% | 66% | 2% | 36% |
| HY&G4&Q2 | 73.0% | 83.5% | 396.0 | 2% | 47% | 3% | 33% | 81.0% | 380.3 | 4% | 48% | 3% | 32% |
| HY&DS&Q2 | 73.0% | 82.5% | 366.2 | 5% | 47% | 6% | 31% | 81.5% | 388.1 | 2% | 46% | 3% | 39% |
| HY&DS&G4&Q2 | 73.0% | 82.0% | 470.9 | 7% | 54% | 7% | 43% | 83.0% | 442.9 | 6% | 54% | 3% | 44% |
| 平均 | **72.5%** | **82.6%** | **404.0** | **4.5%** | **56%** | **4%** | **32%** | **81.6%** | **389.9** | **3.8%** | **56%** | **2.1%** | **36.5%** |

### 3.5　单/多协作模型性能对比分析

综合分析上述单协作与多协作模型场景下ConfiPara方法的实验结果，可以发现引入退出机制的协作方法普遍伴随着一定的推理准确率下降。这一现象主要源于基模型自身存在的固有局限，包括其独立推理时的知识边界限制、生成内容的幻觉倾向，以及在单一路径中自我纠错能力的不足。即便在推理前期通过Token级协作将推理路径引导至正确方向，模型在后续独立推理过程中仍可能受限于其内在能力，最终输出错误结果。该猜想可从实验结果中得到验证。由表4~表6可知，采用全程协作策略的方法（即DuetNet和UniTe）在多数情况下能够达到最优的推理准确率。此外，在基模型能力较弱时，采用固定协作Token数量的DPara-N与UPara-N方法的准确率均出现显著下降，反映出这类方法对基模型性能的敏感性，也进一步印证了上述猜想。

进一步分析可以发现，ConfiPara在单协作模型场景中的推理准确率损失低于多协作模型场景。这一差异表明，多协作模型配置可能引入了额外的性能制约因素，值得深入探究。为解释上述现象，本文提出如下假设：对于给定推理问题，正确的推理路径往往存在多条；然而，每个模型受其自身训练数据与结构限制，仅能覆盖部分可行路径。因此，即便向某一模型提供另一正确路径的部分信息作为推理起点，该模型仍可能因自身局限性而无法独立完成后续正确推理。举例而言，考虑一个问题可以通过“解法1”与“解法2”两种方式求解。若模型A仅掌握“解法1”，则即便在输入中提示其可采用“解法2”，模型A仍难以基于该提示完成有效推理。

在单协作模型场景中，基模型在共识形成过程中通常具备较强的推理路径主导权。这意味着，由于基模型在决策中拥有一定的话语权重，经由Token级协作所达成的“解法”往往落在其擅长的推理范畴之内。因此，若该协作路径正确，基模型在后续独立推理过程中也有较高概率复现正确结果。这一机制解释了为何在单模型协作设定中，整体推理损失相对较低。而在多模型协作中，单个模型的贡献被群体共识稀释，推理链的正确性更依赖于模型间的互补与修正。此时达成的共识解法往往超出任一基模型的固有优势范围，导致其在脱离协作环境后，独立推理准确率更易下降，反映出多模型系统对协作状态更强的依赖性。

尽管如此，在协作模型数量增加时，ConfiPara方法仍能取得优于单模型场景的平均推理准确率。具体而言，基模型DS在ConfiPara-D与ConfiPara-U下的平均推理准确率从单模型场景的67%与69.5%，分别提升至多模型场景的71.3%与75.6%；G4&HY组合的平均推理准确率也从约78.5%与76.5%进一步提高至79.9%与80%。

### 3.6　推理时延分析

本节进一步验证ConfiPara在降低推理时延方面的有效性，全程协作方法与ConfiPara的推理时延比较如图6所示。

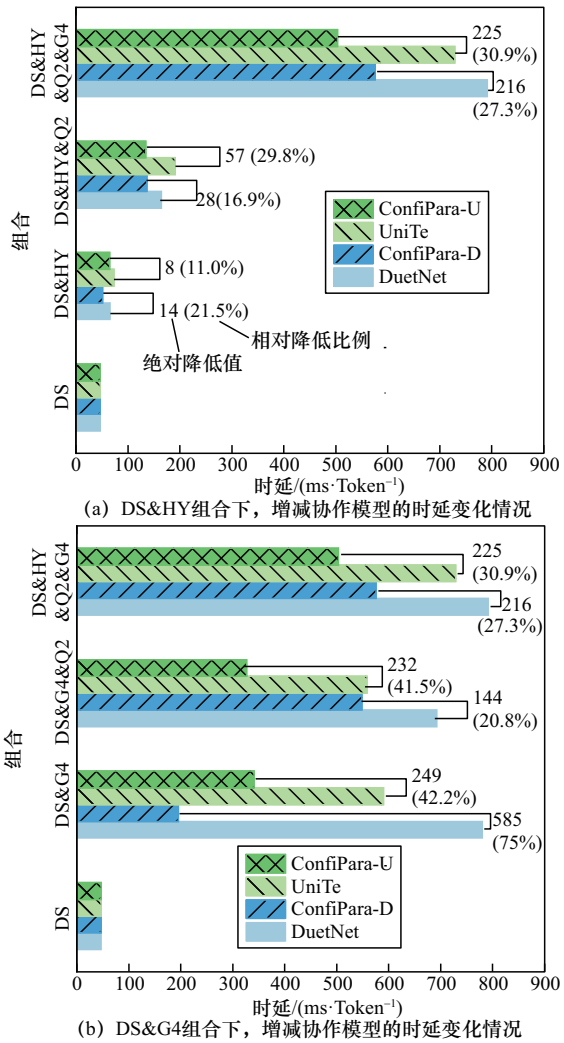



图6　全程协作方法与ConfiPara的推理时延比较

由图6可以看出，随着协作模型数量的增加，所有Token级协作方法的推理时延均呈现上升趋势。这一现象验证了本文的初始判断：Token级协作方法因需同步多方模型的响应，不可避免地会引

入显著的时延开销。在实际部署环境中，单块GPU通常以共享模式服务于多用户任务。当系统负载升高时，每个协作模型所能分配的计算资源相应减少，从而进一步加剧推理时延。这种资源竞争效应会通过协作链路传递并放大，最终导致整体系统时延的显著上升。图6(b)中DS&G4组合在DuetNet方法下的时延异常高于DS&G4&Q2组合，正是GPU资源竞争导致的典型现象。

进一步分析可以发现，在Token级多模型协作场景中，协作模型的选择对系统时延具有决定性影响。DS作为基模型，当DS与HY模型协作时，推理时延可控制在100 ms/Token以内；而与G4模型协作时，时延则显著升高至195~780 ms/Token，二者相差最高达8倍。这一结果揭示出模型互联网中多模型协作的另一个关键权衡问题：除了通过模型选择优化推理准确率外，还需综合考虑协作模型的实际计算能力与通信开销，才能有效控制系统时延、保障用户体验。该发现也为后续研究指明了方向，即需要建立兼顾性能与效率的协作模型选择准则。

最后，在推理时延优化方面，ConfiPara方法展现出显著优势。如图6(a)与图6(b)所示，在不同模型组合下该方法均实现推理时延的显著下降，最大相对降幅达75%。具体而言，由图6(a)可见，与DuetNet相比，ConfiPara-D的推理时延降低了16.9%~27.3%；与UniTe方法相比，ConfiPara-U的推理时延降低了11%~30.9%。特别地，由图6(b)可见，在组合DS&G4中，ConfiPara-D与DuetNet相比，推理时延降低了约75%。这一结果验证了前述论断：当基模型退出协作进行推理后，由于不需要等待协作模型的响应，因此有效降低了推理时延。

### 3.7　消融分析

为验证基模型自信度与信心可靠度的重要性，本节开展了消融实验，结果如图7所示。

由图7可以看出，仅考虑自信度和仅考虑可靠度的消融方法的平均推理准确率都不超过ConfiPara方法。特别地，仅考虑基模型自信度的策略显著劣于完整的ConfiPara方法，平均推理准确率至少下降3%。产生上述结果的原因在于，去除信心可靠度后，基模型更易产生过度自信，从而过早退出协作并转入独立推理，最终引发推理准确率的显著下降；当基模型性能明显劣于协作模型时，这一现象尤为突出。综合上述实验结果，单独依赖信心可靠度或基模型自信度均会使推理准确率降低，验证了二者在ConfiPara方法中的协同与互补作用。

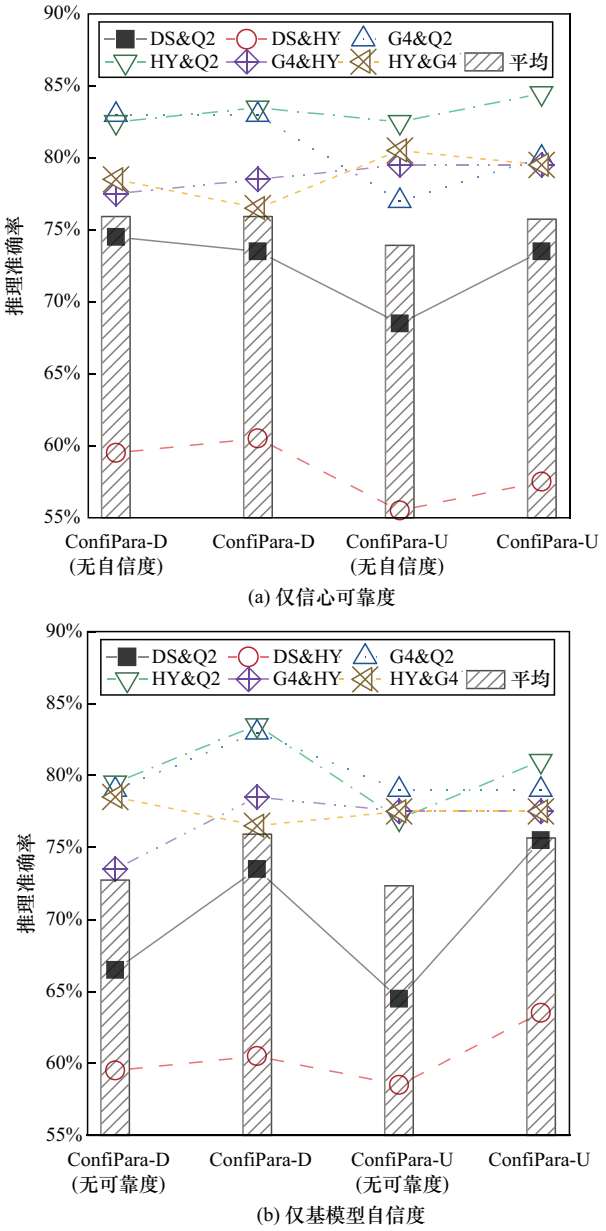



图7　消融实验结果

## 4　结束语

Token级多模型协作推理在模型互联网中具有重要应用潜力，但现有方法采用的全程协作策略带来了较高的Token开销，制约了其实际应用。为应对上述挑战，受自我效能理论的启发，本文设计了一种基于自我效能的Token级多模型协作方法ConfiPara。在ConfiPara中，基模型在高自我效能时退出协作进行独立推理，在保证推理准确率的同时降

低Token开销。实验结果表明，ConfiPara以较小推理准确率损失显著降低了Token开销和推理时延。该研究为模型互联网中的协作推理与通信优化提供了可行路径。在未来的工作中，将聚焦于协作场景下的模型信心评估，进一步优化协作效率。

## 参考文献：


[1] 李哲涛, 曾曦玉, 王建辉, 等. 模型互联网：概念、现状和未来[J]. 计算机研究与发展, 2026, 63: 1-19.
Li Z T, Zeng X Y, Wang J H, et al. AI-model network: concept, current state and future[J]. Journal of Computer Research and Development, 2026, 63: 1-19.

[2] 徐刚, 刘志鹏, 冯骐, 等. 大语言模型在教育信息化中的实践: 规范、框架与应用[J]. 通信学报, 2024, 45(Z2): 229-241.
Xu G, Liu Z P, Feng Q, et al. Practical application of large language models in educational informatics: specification, framework, and applications[J]. Journal on Communications, 2024, 45(Z2): 229-241.

[3] Guo T C, Chen X Y, Wang Y Q, et al. Large language model based multi-agents: a survey of progress and challenges[C]//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. New York: ACM Press, 2024: 8048-8057.

[4] 吴俊儒, 李哲涛, 王建辉, 等. 基于协作关系的模型动态路由[J]. 软件学报, 2025, DOI:10.13328/j.cnki.jos.007498.
Wu J R, Li Z T, Wang J H, et al. Dynamic models routing based on collaborative relationships[J]. Journal of Software, 2025, DOI: 10.13328/j.cnki.jos.007498.

[5] Shen Z J, Lang H, Wang B L, et al. Learning to decode collaboratively with multiple language models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2024: 12974-12990.

[6] Qian C, Li J H, Dang Y F, et al. Iterative experience refinement of software-developing agents[J]. arXiv Preprint, arXiv: 2405.04219, 2024.

[7] Wang Q N, Wang Z H, Su Y, et al. Rethinking the bounds of LLM reasoning: are multi-agent discussions the key? [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2024: 6106-6131.

[8] Kaesberg L B, Becker J, Wahle J P, et al. Voting or consensus? decision-making in multi-agent debate[C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg: ACL, 2025: 11640-11671.

[9] Lu J L, Pang Z L, Xiao M, et al. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models[J]. arXiv Preprint, arXiv: 2407.06089, 2024.

[10] Bandura A. Self-efficacy: toward a unifying theory of behavioral change[J]. Psychological Review, 1977, 84(2): 191-215.

[11] Xu Y, Lu J L, Zhang J J. Bridging the gap between different vocabularies for LLM ensemble[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg: ACL, 2024: 7140-7152.

[12] Huang Y C, Feng X C, Li B H, et al. Ensemble learning for heterogeneous large language models with deep parallel collaboration[C]// Proceedings of the NeurIPS 2024 - The 38th Annual Conference on Neural Information Processing Systems. Vancouver: Curran Associates, 2024: 119838-119860.

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

[14] YAO Y X, WU H, LIU M Y, et al. Determine-then-ensemble: Necessity of top-k union for large language model ensembling[C]// Proceedings of the ICLR 2025 - The 13th International Conference on Learning Representations. Amherst: OpenReview. net, 2025: 101533-101551.

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

[16] Geng J H, Cai F Y, Wang Y X, et al. A survey of confidence estimation and calibration in large language models[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg: ACL, 2024: 6577-6595.

[17] Yin Z Y, Sun Q S, Guo Q P, et al. Do large language models know what they don't know? [C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg: ACL, 2023: 8653-8665.

[18] Zhang M Z, Huang M Q, Shi R D, et al. Calibrating the confidence of large language models by eliciting fidelity[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2024: 2959-2979.

[19] Li Y K, Wang S J, Huang L F, et al. Graph-based confidence calibration for large language models[J]. arXiv Preprint, arXiv: 2411.02454, 2024.

[20] Zhao X R, Zhang H M, Pan X M, et al. Fact-and-reflection (FaR) improves confidence calibration of large language models[C]//Proceedings of the Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL, 2024: 8702-8718.

[21] Zhang W C, Zhang R Q, Guo J F, et al. Pretraining data detection for large language models: a divergence-based calibration method[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2024: 5263-5274.

[22] Yang R X, Rajagopal D, Hayati S A, et al. Confidence calibration and rationalization for llms via multi-agent deliberation[C]// Proceedings of the ICLR 2024 - The ICLR 2024 Workshop on Reliable and Responsible Foundation Models. Amherst: OpenReview.net, 2024: 1-12.

[23] Lin S, Hilton J, Evans O. TruthfulQA: measuring how models mimic human falsehoods[C]//Proceedings of the 60th Annual Meeting of the

Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2022: 3214-3252.

[24] HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understanding[C]// Proceedings of the ICLR 2021 - The 9th International Conference on Learning Representations. Amherst: OpenReview.net, 2021: 1-27.

[25] COBBE K, KOSARAJU V, BAVARIAN M, et al. Training verifiers to solve math word problems[J]. arXiv Preprint, arXiv: 2110.14168, 2021.

[26] Clark C, Lee K, Chang M W, et al. BoolQ: exploring the surprising difficulty of natural yes/No questions[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg: ACL , 2019: 2924-2936.

[作者简介]

**王建辉**（1997– ），男，湖南永州人，暨南大学博士生，主要研究方向为边缘计算、人工智能。

**李哲涛**（1980– ），男，湖南邵阳人，暨南大学教授、博士生导师，主要研究方向为计算机网络、人工智能和安全。

**石伟凡**（2002– ），男，河北邯郸人，暨南大学硕士生，主要研究方向为边缘计算、大模型、人工智能。

**王泽平**（1999– ），男，贵州黔东南人，暨南大学博士生，主要研究方向为算力网络、算力交易和人工智能等。

**郑智润**（1995– ），男，江西上饶人，亚洲大学在站博士后、讲师，主要研究方向为群智系统安全、隐私计算和人工智能安全等。

**李成新**（1996– ），男，湖北黄冈人，湘潭大学博士生，主要研究方向为移动群智感知、数据隐私保护等。