模型互联网中基于输出相似度分簇的模型匹配方法

龙赛琴[1]，赖俊杰[1]，肖勇[1]，刘忠仁[1]，曾丽[2]，李哲涛[1]
（1.暨南大学信息科学技术学院，广东 广州 510632；2.长沙理工大学计算机科学与技术学院，湖南 长沙 410114）

摘要：针对模型互联网中协作模型匹配效率与增益难以兼顾的问题，提出了一种基于输出相似度分簇的模型匹配方法OSC-MM。通过主导–同步模型解耦的相似度度量方法，来刻画模型间的Token级行为差异；利用相似度度量结果进行分簇，并结合代表模型机制，降低匹配复杂度。当新模型接入时，仅需小规模验证，即可完成协作匹配。实验结果表明，OSC-MM与传统实测方法相比，验证开销降低42%–68%；同时，该方法能优先筛选出具有互补性的模型组合，降低同质化并联风险。

关键词：模型互联网；模型匹配；相似度度量；模型协作
中图分类号：TP181
文献标志码：A
doi：10.11959/j.issn.1000

Output Similarity Clustering Based Model Matching Method in AI-Model Network

Long Saiqin[1], Lai Junjie[1], Xiao Yong[1], Liu Zhongren[1], Zeng Li[2], Li Zhetao[1]
1. College of Information Science and Technology, Jinan University, Guangzhou 510632, China
2. School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, 410114，China

Abstract: To address the challenge of balancing efficiency and gain in collaborative model matching within the AI-model network, a method named Output Similarity Clustering based Model Matching (OSC-MM) was proposed. Token-level behavioral differences between models were characterized via a similarity measurement approach that decouples the leader and synchronized models. The resulting similarity measurements were then used for clustering, and a representative model mechanism was introduced to reduce matching complexity. Small-scale validation was employed to complete collaborative matching when a new model joined. Compared with traditional empirical measurement methods, the model matching overhead was reduced by 42%–68% using OSC-MM. Moreover, the method prioritizes complementary model combinations and thereby mitigates the risk of homogeneous parallel deployment.

Key words: ai-model network, model matching, similarity measurement, model collaboration

0 引言

随着大模型技术的快速发展，单一模型在复杂任务中的能力边界逐渐显现。受计算机网络向互联网演进的启发，模型互联网通过连接云、边、端的异构模型，构建协作推理网络，实现跨模型能力整合。在该生态下，用户可将自有模型接入网络，与

收稿日期：2026–06–09；**修回日期：**2026–07–21
通信作者：曾丽，xtuzengli@163.com
基金项目：国家自然科学基金重大科研仪器研制项目（62527823）、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）
Foundation Items： National Natural Science Foundation of China Major Scientific Research Instrument Development Project (62527823), National Natural Science Foundation of China International Cooperation Key Project (W2411053), National Natural Science Foundation of China Joint Fund Key Project (U23B2027), Guangdong Province Basic and Applied Basic Research Major Project (2026B0303000007).

网络中的模型协同完成任务，从而弥补单模型能力不足[1-4]。

在模型互联网中，并联协作作为其中一种重要的协作范式，其允许多个模型在推理阶段同时参与，并通过融合各模型的预测结果来提升整体性能[2,5]。其中，Token级聚合是一种兼顾效率与性能的实现方式，无需复杂的语义对齐，仅在输出分布层面进行融合，能有效融合多个来源的分布信息，抑制单一模型的噪声与偏差，避免语义对齐带来的失真。然而并联协作能否产生有效增益，很大程度上取决于参与模型之间是否能够提供差异化信息：若多个模型在预测行为上高度一致，无论聚合策略多精细，都难以引入额外信息增益[6-7]。

现有研究主要关注聚合策略设计，而对模型匹配问题缺乏系统性研究。在实际部署中，协作模型的选择通常依赖人工经验或简单规则，缺乏系统化方法。这一研究空白带来了两个直接后果：一方面，模型匹配效率低下，难以适应模型互联网中的大规模模型网络；另一方面，现有并联协作研究往往隐含假设所选模型具有较强互补性，或直接采用随机组合，未充分考虑模型间相似性对协作效果的影响，从而可能高估并联协作的实际收益，难以反映模型互联网中“异构性与同质化并存”的真实复杂环境。

然而，当前主流模型普遍基于高度重叠的预训练数据，并采用相似的训练范式，使得不同模型在知识来源与决策偏好上呈现出明显的同质化特征[8-10]。在缺乏有效筛选机制的情况下，模型之间往往存在信息冗余，限制了并联协作的实际效果。因此，在实际应用中，模型互联网面临一个关键瓶颈：如何从大规模候选模型中高效匹配合适的协作对象。

在模型匹配过程中，关键环节便是刻画模型输出行为、量化模型之间的相似性，从而识别出能提供差异化信息的协作对象。此外，在大规模模型网络中，对所有可能的模型组合进行逐一判断不仅计算上不可行，也因缺乏结构化组织而难以实现高效检索。

为应对上述挑战，本文提出一种基于输出相似度分簇的模型匹配方法（output similarity clustering based model matching, OSC-MM）。该方法基于模型输出行为的相似性对模型进行分簇，并通过代表模型构建紧凑的簇级结构。在新模型接入时，仅需与代表模型匹配即可确定其归属，并在跨簇范围内筛选候选协作模型，从而避免同质化组合并降低验证开销。最终通过并联验证确定较好的协作对象，实现效率与性能的平衡。

本文的主要贡献如下：

1)针对模型互联网中协作模型匹配效率低下与组合同质化的问题，本文提出一种基于输出相似度分簇的模型匹配方法OSC-MM，通过对模型空间进行分簇建模，并在簇级别完成候选筛选。在保证模型差异性的同时，显著降低协作模型搜索开销。

2)提出一种面向并联协作的输出相似度度量方法，通过主导－同步模型划分来解耦生成过程，在不干扰生成轨迹的前提下刻画模型间的行为差异，从而获得相似度度量结果。

3)设计了一种基于代表模型的高效分簇与增量接入机制，将新模型匹配复杂度由与模型规模相关降低为与簇数量相关，实现大规模模型网络中的高效扩展。

4)实验结果表明，所提方法通过离线相似度计算替代大量在线并联验证，显著降低了协作验证成本。与传统实测方法相比，本文方法可使模型匹配验证开销降低42%~68%，且模型规模越大，降本效果越明显。同时与其他方法相比，能在实验设置的两种环境下稳定筛选出具有互补性的模型组合，有效提升并联协作的匹配效率。

1 相关工作

1.1 模型相似性度量

大语言模型因数据重叠、知识蒸馏传递及对齐收敛等因素，其输出行为表现出显著趋同[11-13]。针对这一现象的量化评估，研究者提出了多种相似性度量方法。根据度量与生成过程是否解耦，现有工作可大致划分为独立生成后比较与协作过程中同步度量两类。

独立生成后比较方法最为直接。基于生成文本的语义度量，通过比较两个模型对同一输入的完整回复来计算语义重叠程度，典型指标包括BLEU、ROUGE等n-gram重叠指标，以及基于上下文嵌入的BERTScore[14]。Jiang[11]等人进一步构建了Infinity-Chat数据集，借助语义聚类量化模型间的同质性。基于输出概率分布的度量主要以分支因子和多

重 Token 散度[15]为代表，其中，Yang[13]等人利用分支因子衡量生成过程中合理候选 Token 的数量，依此反映模型输出的确定性。这类方法的优点是独立于生成过程，不干扰模型自身的输出；其不足之处在于无法反映模型在协作过程中的相似度，且难以对齐生成长度不一的模型输出。

协作过程中同步度量方法常见于 Token 级多模型协作框架。在典型的 Token 级协作框架（如 DuetNet[2]、EVA[16]、UniTe[17]）中，多个模型基于完全相同的上文并行预测下一词集的概率分布，随后通过投票、加权或动态选择等聚合策略确定最终的输出结果。在此类框架下，Yao 等人[17]通过比较各模型对 Top-K 候选 Token 的排序一致性来衡量预测差异。然而，这类方法存在根本性局限：度量过程与生成过程高度耦合。由于所有模型始终共享相同的上文，且每步聚合得到的 Token 会作为下一轮生成时的公共输入。因此，随着生成步数增加，模型间的相互影响不断累积，促使各自的预测分布逐渐趋同。最终导致输出内容趋于一致。

针对上述问题，本文方法实现了相似性度量与生成过程的解耦，实现 Token 级的度量，为多模型协作中的相似性度量提供了可靠基础。

1.2　模型匹配

在模型互联网中，并联协作的有效性依赖参与模型的能力和互补性。如何为用户自有模型匹配最合适的协作模型，即模型匹配问题，尚未得到充分研究。

从更广泛的视角看，模型选择领域已有丰富的研究积累[18]，现有工作主要分为两类。基于性能的选择[19-20]根据模型在基准测试上的准确率、推理速度或成本进行排序，直接选取综合表现最优的模型。这类方法简单有效，但忽略了模型之间的差异性，容易导致多个高性能但高度相似的模型被同时选中，协作增益有限。基于任务特征的选择[21-23]则训练一个路由网络，根据输入查询的领域或难度，将查询动态分配给最擅长的模型。这类方法以单模型推理为目标，而非为已有模型寻找协作对象。

在多模型集成场景中，Cruz 等人[24]实现动态集成选择策略，根据模型在局部区域内的性能或置信度动态筛选子集。尽管它涉及多个模型的组合，但其筛选依据仍是各模型的绝对表现（如局部准确率或预测置信度），而非模型之间的相对差异性。因此，该方法并不能直接用于解决模型匹配问题。

总体而言，现有工作主要侧重于从候选模型中选出一个最优模型来独立完成任务，而忽略了另一种场景：用户将自有模型接入模型网络后，需要从海量候选模型中为其匹配一个协作模型，从而能在弥补自有模型短板的同时，提升并联协作的多样性收益。

针对上述问题，本文提出一种基于输出相似度分簇的模型匹配方法，能够在保证模型差异性的同时，显著降低协作模型搜索开销。

2　OSC-MM 方法

2.1　系统模型与问题定义

模型互联网中的协作模型匹配架构如图 1 所示，其核心由模型互联网平台、模型实体与用户 3 个部分构成。平台作为中枢，负责模型的接入管理与匹配调度，用户则是任务的发起者与自有模型的提供者。

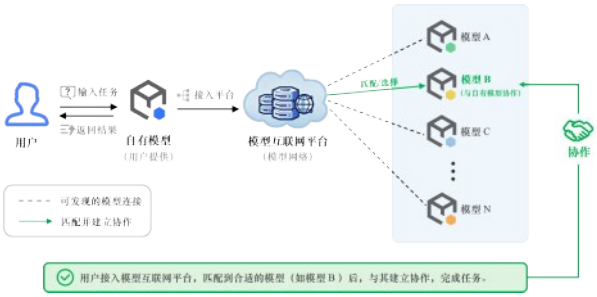

图 1　模型互联网中协作模型匹配架构

在上述架构中，用户不仅是任务的发起方，还可将自有模型接入平台。具体执行过程如下：首先，用户将原始任务提交给自有模型；其次，该模型作为主导方接入平台，发起协作请求；随后，平台根据匹配策略为其筛选合适的协作模型；最后，自有模型与协作模型通过网络交互，协同处理任务。协作模型仅返回其计算结果。

本文要解决的核心问题是：当平台中有大量候选模型时，如何为新加入或发起任务的自有模型，高效地匹配出较好的协作模型。

2.2　方法总体框架

本文方法的总体流程如图 2 所示，OSC-MM 方法将协作模型的匹配过程分为两个阶段：新模型加入前的模型网络构建，与新模型加入后的快速接入匹配。其核心思想是：利用模型间的相似度度量来

刻画模型间的输出行为，再通过聚类预先构建一个模型关系网络；待新模型加入时，将匹配过程分解为“簇级定位”与“候选验证”两步，将全量搜索转化为跨簇搜索。

2.2.1　模型网络构建

聚类分簇：计算模型池中所有模型对之间的输出相似度，并以此构建相似度矩阵。随后，基于该矩阵对模型进行聚类，将其划分为若干相近的簇。

结构化表示：针对每个簇，从中提取两类关键模型：其一为代表模型，用于概括该簇的整体特征；其二为Top模型，作为该簇内优先推荐的协作候选。最终构建出一个结构化的模型网络。

2.2.2　快速接入与匹配

簇级定位：当新模型加入时，仅计算其与各簇代表模型的相似度，将其归入行为最相似的簇。

候选验证：将新模型与其他簇中的Top协作模型进行并联实测，根据实际性能增益，选出最优的协作模型。

通过上述设计，OSC-MM方法将全模型间的并联验证规模，压缩为仅与少量候选模型之间进行实测，从而将原来的全模型搜索转化为分层搜索，有效降低了相似度计算与并联验证的整体开销。

2.3　基于主导-同步模型划分的相似度度量方法

在多模型并联协作过程中，若简单平均聚合各模型的输出概率，生成过程会倾向于模型间的共性分布，导致表达趋同。它会干扰我们对模型间输出行为的客观评估。为解决这一问题，本节提出一种主导-同步模型角色划分的模型输出相似度度量方法。如图3所示。

设模型集合为 $\mathcal{M}=\{M_1,M_2,\cdots,M_k\}$，对于任意模型对 (M_i,M_j)，指定 M_i 为主导模型，M_j 为同步模型。在时间步 t 时，给定当前上下文 $C_t=(x_1,x_2,\cdots,x_{t-1})$，主导模型与同步模型分别预测下一个词元Top-k概率分布：

$$\mathbf{p}_i^{(t)}=P_{M_i}(\cdot\mid C_t),\mathbf{p}_j^{(t)}=P_{M_j}(\cdot\mid C_t). \tag{1}$$

其中 $\mathbf{p}_i^{(t)}$ 与 $\mathbf{p}_j^{(t)}$ 分别表示 M_i 和 M_j 在时间步 t 预测的下一词元Top-k概率分布。在协作过程中，主导模型依据其概率分布生成实际词元加入上下文：

$$x_t=\arg\max \mathbf{p}_i^{(t)},\quad C_{t+1}=C_t\oplus x_t. \tag{2}$$

在此基础上，定义两者在时间步 t 的输出相似度为余弦相似度：

$$sim_{i,j}^{(t)}=\frac{<\mathbf{p}_i^{(t)},\mathbf{p}_j^{(t)}>}{\|\mathbf{p}_i^{(t)}\|_2\cdot\|\mathbf{p}_j^{(t)}\|_2}. \tag{3}$$

整个生成序列上的模型输出相似度，即为各时间步输出相似度的平均值：

$$Sim(M_i,M_j)=\frac{1}{T}\sum_{t=1}^{T}sim_{i,j}^{(t)}. \tag{4}$$

其中，T 表示生成序列的长度。

在该方法中，同步模型的输出仅用于相似度计算，而不参与上下文更新；上下文完全由主导模型生成的结果逐词递推。因此，本方法能够在避免聚合策略干扰生成轨迹的前提下，衡量不同模型在词

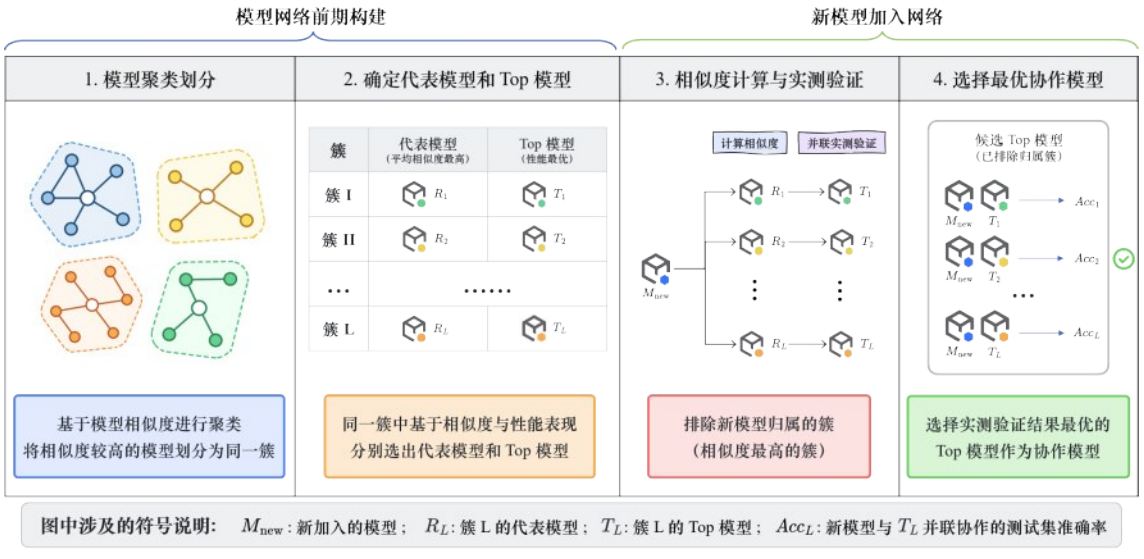

图2　OSC-MM方法中匹配协作模型的整体流程

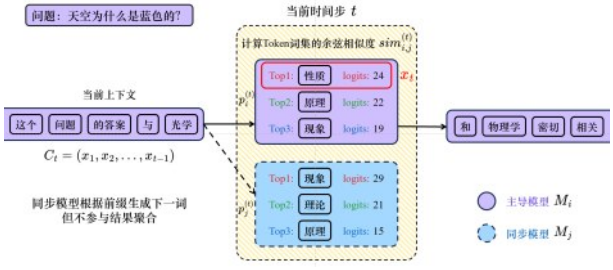

图3　主导-同步模型角色划分的输出相似度计算示意图

元级生成行为上的相似性。

通过遍历所有模型对(M_i,M_j)并重复上述过程，可得到完整的模型输出相似度矩阵。在实际实现中，可预先通过主导模型生成完整的上下文序列$\{C_t\}_{t=1}^{T}$，并在相似度计算阶段复用该序列作为输入提供给各模型进行前向预测，从而无需为每个同步模型重新生成上下文，有效降低计算开销。此外，由于同步模型仅执行条件概率分布的前向计算，该过程可通过批处理方式并行实现，能够进一步提升相似度的计算效率。

算法 1　基于主导-同步模型划分的模型输出相似度度量方法

输入 模型集合$\mathcal{M}$；初始上下文C_0；生成长度T

输出 模型输出相似度矩阵$S\in\mathbb{R}^{k\times k}$

1)初始化相似度矩阵$S\leftarrow 0^{k\times k}$

2)**for** $i=1$ to k **do**

3) 令M_i为主导模型

4) 初始化上下文序列$\mathcal{C}\leftarrow\varnothing$, $C\leftarrow C_0$

5) **for** $t=1$ to T **do**

6) 保存当前上下文$\mathcal{C}\leftarrow\mathcal{C}\cup\{C\}$

7) 主导模型输出分布$\mathbf{p}_i^{(t)}$

8) 生成下一词元并更新上下文

$$x_t=\arg\max\mathbf{p}_i^{(t)},\ C\leftarrow C\oplus x_t$$

9) **end for**

10) **for** $j=1$ to k **do**

11) 令M_j为同步模型，相似度$s\leftarrow 0$

12) **for** $t=1$ to T **do**

13) 取主导模型生成轨迹$C_t\in\mathcal{C}$

14) 计算同步模型输出分布$\mathbf{p}_j^{(t)}$

15) 计算时间步t分布相似度$sim_{i,j}^{(t)}$

16) $s\leftarrow s+sim_{i,j}^{(t)}$

17) **end for**

18) $S_{i,j}\leftarrow\dfrac{s}{T}$

19) **end for**

20)**end for**

21)**return S**

2.4　基于分簇的层次化模型匹配

为降低相似度计算与协作模型匹配的整体开销，本节提出一种基于分簇的中心化相似度建模与分层筛选方法。该方法首先对模型进行相似度驱动的分簇划分，并在簇内选取代表模型与Top模型；随后，对于新加入的模型，仅在簇级别进行相似度计算与协作模型筛选，并通过有限的并联实测验证确定最优协作模型，从而将原有的全模型搜索问题转化为簇级搜索问题，显著减少整个过程中的计算开销。

2.4.1　模型分簇与簇内代表模型选取

基于模型输出相似度矩阵，将模型集合$\mathcal{M}$划分为L个互不相交的簇：

$$\mathcal{M}=\bigcup_{\ell=1}^{L}\mathcal{C}_\ell,\qquad \mathcal{C}_\ell\cap\mathcal{C}_{\ell'}=\varnothing(\ell\neq\ell').\tag{5}$$

其中，$\mathcal{C}_\ell$表示第ℓ个模型簇，通常满足$L\ll k$。

对于每个簇$\mathcal{C}_\ell$，定义代表模型$M_\ell^c\in\mathcal{C}_\ell$。选取与簇内其他模型平均相似度最高的模型来确定：

$$M_\ell^c=\arg\max_{M_i\in\mathcal{C}_\ell}\frac{1}{|\mathcal{C}_\ell|-1}\sum_{\substack{M_j\in\mathcal{C}_\ell\\ j\neq i}}Sim(M_i,M_j).\tag{6}$$

代表模型刻画了簇的整体输出行为特征，实现对模型空间的压缩表示。

2.4.2　中心化相似度计算与归属判定

对于新模型M_{new}，计算其与各簇代表模型M_ℓ^c的相似度：

$$\ell^*=\arg\max_{\ell\in\{1,\cdots,L\}}Sim(M_{\text{new}},M_\ell^c).\tag{7}$$

当$Sim(M_{\text{new}},M_{\ell^*}^c)\geqslant\tau$时，则将$M_{\text{new}}$归入簇$\mathcal{C}_{\ell^*}$；否则，将其视为新的独立簇。阈值$\tau$用于控制簇内紧凑性与簇间区分性。

该策略将原本“对所有模型两两比较”的过程转化为“对代表模型的选择性比较”，使得增量场景下单个模型的计算复杂度由$O(K)$降低至$O(L)$。

2.4.3　基于Top模型的协作模型匹配

在确定新模型的簇归属后，为进一步选择其最优协作对象，本文引入Top模型进行候选筛选。

对于每个簇$\mathcal{C}_\ell$，根据模型在测试集$\mathcal{D}_{\text{test}}$上的性

能表现，选取最优模型作为该簇的 Top 模型：

$$M_{\ell}^{top} = \arg\max_{M_i \in \mathcal{C}_{\ell}} Acc(M_i, \mathcal{D}_{\text{test}}). \quad (8)$$

其中，$Acc(M_i, \mathcal{D}_{\text{test}})$ 表示模型 M_i 在测试集 $\mathcal{D}_{\text{test}}$ 上的准确率。

由于同一簇内模型具有较高相似性，其并联协作往往缺乏有效互补性。因此，在协作模型匹配阶段，本文排除新模型所属簇 $\mathcal{C}_{\ell^*}$ 仅在其他簇中进行候选搜索。对于每个候选簇 $\mathcal{C}_{\ell}(\ell \neq \ell^*)$，计算新模型 M_{new} 与该簇的 Top 模型 M_{ℓ}^{top} 进行并联推理（通过词元聚合机制）。最终，选择能够取得最高并联验证性能的 M_{ℓ}^{top} 作为协作对象：

$$M^* = \arg\max_{M_{\ell}^{top}, \ell \neq \ell^*} Acc\left(\{M_{\text{new}}, M_{\ell}^{\text{top}}\}, \mathcal{D}_{\text{test}}\right). \quad (9)$$

通过上述方式，传统方法需要将新模型与网络中所有已有模型进行并联实测验证，其复杂度为 $O(K)$；而本文方法仅需与各簇 Top 模型进行验证，复杂度降低为 $O(L)$。由于通常满足 $L \ll K$，该方法能够显著降低协作模型匹配的验证成本，同时保留跨簇模型之间的互补性搜索能力。

算法 2　基于分簇的中心化相似度建模与协作模型匹配

输入 模型集合 $\mathcal{M}$；测试集 D_{test}；验证集 D_{val}；阈值 τ；新模型 M_{new}

输出 协作模型 M^*

1)// 阶段 1：模型分簇

2)将 $\mathcal{M}$ 划分为 L 个簇 $\{C_1, C_2, \cdots, C_L\}$

3)**for** $\ell = 1$ to L **do**

4) 选取代表模型 $M_{\ell}^{\mathcal{C}}$

5) 选取 Top 模型 M_{ℓ}^{top}

6)**end for**

7)// 阶段 2：簇归属判定

8)初始化 $s_{max} \leftarrow -\infty$, $\ell^* \leftarrow -1$

9)**for** $\ell = 1$ to L **do**

10) $s_{\ell} \leftarrow Sim(M_{\text{new}}, M_{\ell}^{\mathcal{C}})$

11) **if** $s_{\ell} > s_{max}$ **then**

12) $s_{max} \leftarrow s_{\ell}$, $\ell^* \leftarrow \ell$

13) **end if**

14)**end for**

15)**if** $s_{max} < \tau$ **then**

16) 创建新簇 $M_{\text{new}} \subset C_{L+1}$

17)**end if**

18)// 阶段 3：协作模型匹配

19)初始化 $A_{max} \leftarrow -\infty$，最优模型 $M^* \leftarrow \varnothing$

20)**for** $\ell = 1$ to L **do**

21) **if** $\ell \neq \ell^*$ **then**

22) $A_{\ell} \leftarrow \text{Acc}(\{\text{M}_{\text{new}}, M_{\ell}^{top}\}, \text{D}_{\text{val}})$

23) **if** $A_{\ell} > A_{max}$ **then**

24) $A_{max} \leftarrow A_{\ell}$, $M^* \leftarrow M_{\ell}^{top}$

25) **end if**

26) **end if**

27)**end for**

28)**return** M^*

3　实验结果分析

3.1　实验设置

本文所使用的实验数据集如表 1 所示。针对每个数据集，随机抽取 100 – 200 道样本构建统一测试集。为保证实验的公平性与可比性，所有模型均采用相同的提示词模板进行推理，模型并联协作中的聚合方式采用简单平均聚合，即各自输出概率分布等权求平均后，作为最终预测结果。

如表 2 所示，本文共选取 12 个开源大模型，其中前 10 个作为模型网络中的基础模型，后 2 个作为新接入的测试模型。实验环境由一台高性能服务器与多台边缘设备构成：服务器端配备 4 块 A100 GPU(80GB)，边缘设备包括 RTX 3090、RTX 4090 以及 Jetson 64GB 等不同算力设备，以模拟模型互联网中的异构计算环境。

3.2　参数影响分析

3.2.1　Top-k 参数影响分析与取值选择

本节研究 Top-k 参数对模型输出相似度的影响，实验选取上述 10 个基础模型，在 Chinese- Cosmopedia 数据集上进行。设置 Top-k = [3, 5, 7, 10, 15, 20] 共六组取值，覆盖从较小到较大的采样范围，以观察不同采样设置下的相似度变化。

为降低随机性因素的干扰，实验中统一将模型温度及其他相关随机参数设置为较低水平，使模型输出尽可能稳定，从而凸显 Top-k 参数对输出相似度的独立作用。

图 4 展示了不同 Top-k 取值下模型输出相似度的热力图分布。图中横坐标对应同步模型，纵坐标对应主导模型，颜色深浅表示两者的输出相似度，颜色越深表示相似度越高。从整体上看，相似度矩

表 1　　实验选取的实验数据集及对应的提示词模板

数据集	题目类型	提示词模板
CEVAL[25]	中文选择题	请回答下面中文选择题：**{question}**\n 用 2-3 句话写出解题关键，最后一行输出“答案：X”。要求输出分析过程和答案。
CMMLU[26]	中文选择题	请回答下面中文选择题：**{question}**\n 用 2-3 句话写出解题关键，最后一行输出“答案：X”。要求输出分析过程和答案。
MMLU[27]	英文选择题	Please answer the following English multiple-choice question: **{question}**\n Write 2-3 sentences explaining the key reasoning, then output "Answer: X" on a new line. Only output the reasoning and the answer.
GSM8K[28]	数学推理题	Please answer the following math question: **{question}**\n Solve the math problem step by step. End with "#### {answer}".
Chinese Cosmopedia[29]	开放性问题	请用中文回答以下问题：**{question}**\n

表 2　　实验选取的 12 个开源大模型

简称	模型	发布者
glm4	glm-4-9b	智谱 AI
phi4	phi-4-mini-3.8b	Microsoft
kimi	kimi-vl-a3b-2.8b	月之暗面
la31	llama-3.1-8b	Meta
la32	llama-3.2-3b	Meta
q253	qwen2.5-3b	阿里巴巴
q257	qwen2.5-7b	阿里巴巴
q304	qwen3-4b	阿里巴巴
g304	gemma3-4b	Google
g327	gemma3-27b	Google
g405	gemma-4-5.1b	Google
mi03	ministral-3b	Mistral AI

阵呈现出近似对称的结构（详细分析在 3.2.2 展开），不同 Top k 参数下的颜色分布差异很小。

为进一步观察变化趋势，图 5 选取其中 6 组具有代表性、覆盖不同相似度水平的模型组合进行重点分析。

实验结果表明，随着 Top-k 的增大，模型输出相似度整体呈现下降趋势，且下降幅度逐渐趋于平缓。

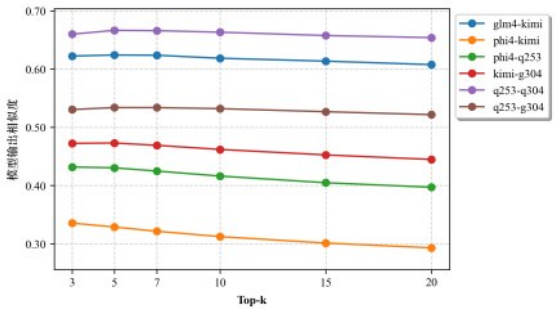

图 5　不同模型组合输出相似度随 Top-k 变化的趋势

因此，综合考虑候选词集合大小与输出相似度之间的关系，本文在后续实验中将 Top-k 固定为 10，在保证生成多样性的同时维持较高的一致性水平。

3.2.2　相似度矩阵对称性分析

为分析不同角色配置（即主导模型与同步模型角色互换）对相似度结果的影响，本节在上述模型输出相似度矩阵的基础上，开展相似度矩阵的对称性分析实验。通过对比不同角色配置下的结果，对模型输出相似度计算的一致性与对称性进行评估。

表 3　　相似度矩阵对称性分析结果

对称相关系数	平均对称绝对误差	最大对称绝对误差
99.93%	0.006	0.016

表 3 给出了相似度矩阵的对称性分析结果。观察发现，对称相关系数接近 1，说明矩阵在整体结构上具有较高一致性；平均对称绝对误差较小，表明整体非对称程度较低；最大对称绝对误差也处于较小范围，说明不存在明显的局部异常偏差。

从实验结果来看，相似度矩阵呈现出一定程度的非对称性。该差异主要来源于不同模型在分词策略与生成模板上的不一致性。例如，模型 A 生成“你好”时，通常采用逐步生成的方式，依次输出“你”及其后续字符；而模型 B 则可能以整体生成的方式一次性输出完整结果。

上述差异会对生成过程产生不同影响：当模型 A 作为主导模型时，其生成前缀中已包含“你”，可能引导模型 B 在后续生成中进行语义补全或扩展，从而引入更多不确定性；而当模型 B 作为主导

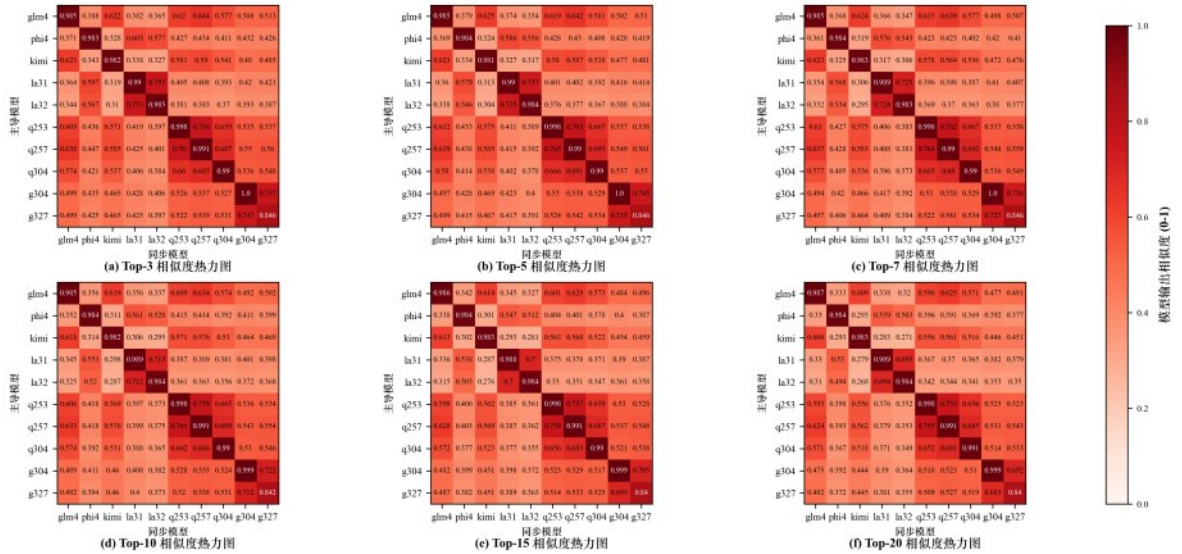

图4　不同 Top-k 下模型输出相似度的热力图分布

模型时，由于其直接生成完整表达“你好”，对后续模型的语义约束更加明确，对生成过程的干扰相对较小。

由于分词策略与生成模板上的不一致，不同角色配置下的生成路径存在差异，进而导致相似度计算结果出现一定程度的不一致性。尽管如此，从整体结果来看，相似度矩阵仍表现出近似对称的结构，误差水平较低，说明不同角色配置对整体相似度分析结果的影响较为有限。

3.2.3　分簇阈值的影响分析与取值选择

为确定合理的分簇阈值，本节基于上述十个基础模型构成的网络，对不同阈值 τ 下的聚类结果进行了系统分析。具体而言，通过设置不同阈值并基于相似度进行模型划分，获得对应的分簇结构。

对于每一组阈值，分别统计分簇数量、簇内平均相似度、簇间平均相似度及偏离程度等指标，从紧凑性与区分性两个方面对分簇结果进行综合评估。同时，通过比较不同阈值下各项指标的变化趋势，分析分簇结果的稳定性。

图6展示了不同分簇数量条件下的簇内平均相似度、簇间平均相似度及偏离程度的变化情况。整体来看，随着分簇数量的增加，各项指标呈现逐步上升趋势。对不同分簇结果进行定量比较发现，当分簇数量为3时，簇内紧凑性与簇间区分性已达到较好水平；进一步增加至4个簇时，偏离程度的改善十分有限。相比之下，将10个模型划分为3个簇，能够在保证各簇具备足够规模的前提下，有效

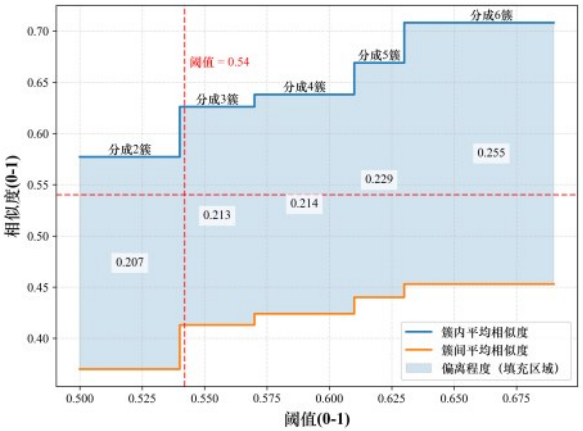

图6　不同分簇数量下的聚类质量评估指标变化趋势

刻画模型间的行为差异，避免因划分过细导致的簇结构不稳定与可解释性降低。

因此，综合考虑分离效果、簇规模的合理性以及结果的可解释性，本文选择将模型集合划分为3簇，并选取分簇阈值 $\tau = 0.54$ 作为后续分析与实验的基础。该阈值设定兼顾两方面需求：一是确保当前十个基础模型恰好能够划分为三个簇，以获得结构清晰、可解释的初始分簇；二是保证后续接入的新模型有较大概率被归入已有三个簇中的某一簇，从而维持分簇结构的稳定性与可扩展性。

该结果同时表明，即便在较小规模的模型集合中，模型间的相似度分布仍呈现出明显的聚集特性，验证了采用分层结构建模的合理性。

3.3　模型并联性能影响因素分析

本节通过实验分析模型个体能力与模型间相似度对并联性能的影响，旨在验证OSC-MM方法中两个关键设计的前提假设：第一，协作增益离不开单模型能力支撑，但并不完全由其决定；第二，跨簇选择策略的合理性，即低相似度组合的模型协作收益更高。

依据模型间相似度进行筛选：选取相似度高于阈值的组合作为高相似度组，低于阈值的组合作为低相似度组，在此基础上对比分析两类组合在并联结构下的性能提升效果。其中，相似度阈值根据前文分簇实验的结果确定（$\tau = 0.54$），该阈值能够较好地区分模型间的相似性差异，确保两组样本在相似度上具有明显区分度。

3.3.1　模型个体能力对并联性能的影响分析

为分析模型个体能力对并联性能的作用，本小节统计了各模型组合中单模型的平均与最大能力表现，并采用皮尔逊相关系数，对这两项指标与并联性能之间的关联进行定量刻画。图7展示了这两项指标与并联性能的相关性。

从图7中可以看出，并联性能与单模型平均能力及最大能力均呈显著正相关，相关系数均达到0.93。这表明并联性能在很大程度上受限于组合中模型的整体水平，单模型能力越强，并联后的表现通常也越好。

结合散点分布可以发现，大部分数据点分布在回归直线附近，说明该相关关系具有较好的稳定性。然而，仍有部分样本偏离趋势线。这表明除模型能力外，并联性能还可能受到模型间相似性或差异性的影响。

针对这一现象，本文在后续实验中对相关样本进行深入分析，重点从模型间相似性与差异性角度探究其对并联结果的影响机制。

3.3.2　模型间相似度对并联性能的影响分析

为进一步验证OSC-MM方法跨簇匹配策略的合理性，本节以“并联性能相较于组合平均性能的提升幅度”为指标，分析模型间相似度对协作增益的影响，结果如图8所示。

从图8中可以看出，低相似度组合的性能提升整体显著高于高相似度组合。低相似度组合的中位数明显更高，表明低相似度组合在协作时更可能获得稳定且显著的性能增益。同时，低相似度组合的

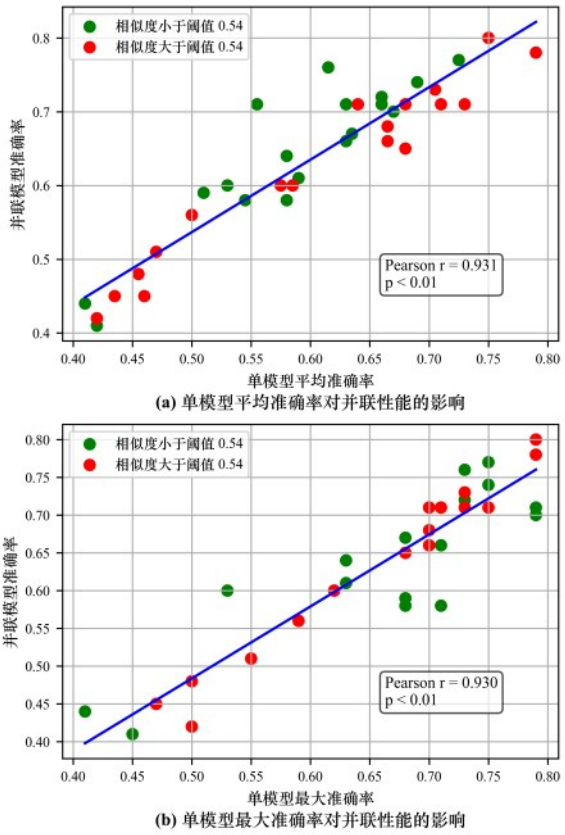

图7　单模型个体能力对并联性能的影响

箱体范围更大，说明不同组合之间的性能提升存在较大差异，即结果具有一定的波动性。另外，低相似度组合中存在若干高增益的离群点，说明当模型间具有较强的互补性时，并联结构能够有效突破单模型能力的限制。

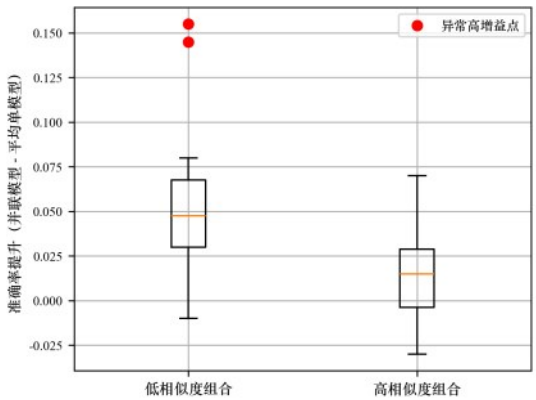

图8　不同相似度组合下的并联性能提升幅度对比

相比之下，高相似度组合的性能提升整体有限，中位数更接近于零，箱体分布也更加集中。这表明当模型输出行为高度相似时，各模型提供的信

息存在较高冗余，并联过程难以引入新的有效信息，增益空间较小。

综上所述，模型个体能力奠定了并联性能的基础，模型输出相似度是影响协作增益大小的关键因素。这一结论验证了OSC-MM方法的核心设计：跨簇匹配能够有效利用模型间差异性，带来更显著的协作增益。

3.4 簇内选优策略分析

为验证OSC-MM方法中“簇内选优”这一策略的合理性，本节设计了簇内对比实验。

实验在每个簇内选取能力排名前两位的模型，即簇内最优模型与次优模型，分别与固定的用户模型进行并联推理，对比其协作性能。实验结果如表4所示。

表4 簇内最优与次优模型协作性能对比

用户模型	类别	簇I	准确率	簇II	准确率	簇III	准确率
mi03	最优	q257	74.80%	g327	62.30%	la31	57.60%
	次优	q304	73.20%	g304	48.10%	la32	54.60%
g405	最优	q257	74.70%	g327	61.60%	la31	54.10%
	次优	q304	71.80%	g304	50.50%	la32	51.00%

从表4中可以观察到，在不同用户模型和不同簇的设置下，选择簇内最优模型进行并联，其性能表现一致且稳定地优于次优模型。以簇II中用户模型mi03为例，与最优模型g327并联的性能为62.30%，而与次优模型g304并联仅为48.10%，性能下降幅度显著。

这一现象验证了OSC-MM方法的前提，即在簇内模型行为高度相似、信息重叠程度较高的条件下，模型间的协作贡献差异主要体现在能力水平上。此时，更强的单模型能力通常能更有效地整合与提炼冗余信息，带来更优的协作效果。

基于上述分析，OSC-MM方法所采用的“簇内选优”策略是有效的：在信息高度冗余的簇内，直接选取能力最优的模型，能够在不显著损失协作潜力的前提下，稳定地逼近该簇所能提供的最佳协作效果。相较于在簇内对所有模型进行遍历式并联验证，OSC-MM方法在保证协作性能的同时，进一步降低了匹配阶段的计算开销，提升了整体方法的可扩展性。

3.5 对比实验

为验证OSC-MM方法在用户接入场景下的有效性，本节设计了协作模型匹配的对比实验。实验选取两个具有不同特征的用户模型，以模拟模型互联网中可能出现的不同分布情况：用户模型g405在候选池中存在多个与其行为高度相似的模型，而用户模型mi03在候选池中则无明显相似模型。该设置旨在检验匹配策略在不同相似性结构下的适应性与鲁棒性。

实验固定用户模型与候选模型池，在统一的Token级并联推理框架下，仅改变协作模型的选择策略，以对比不同匹配方法对并联结果准确率的影响。本文选取以下七种代表性策略进行比较：

本文方法（OSC-MM）：综合考虑候选模型的个体能力以及其与用户模型之间的模型输出相似度，通过分簇与跨簇筛选，选取能力强且具备差异性的协作模型。

能力优先：直接选择候选池中单模型性能最优者作为协作对象，模拟仅以能力为导向的匹配策略。

高相似度优先：选择与用户模型相似度最高的模型进行协作，用以分析高同质性组合的效果。

低相似度优先：选择与用户模型相似度最低的模型进行协作，用以分析高差异性组合的效果。

随机选择：从候选池中随机选择一个模型作为协作对象，代表无导向的基线匹配策略（测试4组取平均），用于衡量其他方法的相对增益。

标签多样性：在候选模型池中优先挑选与用户自有模型能力标签形成互补的模型，从擅长任务领域的维度去进行分类。

贪心错误多样性策略（greedy error diversity, GED）[10]：基于Yule's Q统计量量化模型在验证集上的错误相关度，并采用贪心机制优先选取与用户模型错误相关性最低的候选模型作为协作对象。

实测表现优先（经验选择）：将候选池中所有模型逐一与用户模型并联，基于验证集表现选择最优协作模型。该方法代表经验选择，但需付出高昂的验证成本。

对于上述策略，均构造由用户模型与所选协作模型组成的并联组合，采用简单平均聚合的方式，并在相同测试、验证集上进行评估。

表5展示了不同匹配策略在多个数据集上的并

表5　　对比实验结果

方法	数据集/新增模型								综合
	CEVAL		CMMLU		MMLU		GSM8K		
	mi03	g405	mi03	g405	mi03	g405	mi03	g405	
能力优先	0.77	0.76	0.68	0.63	**0.77**	0.725	0.73	0.67	0.717
高相似度优先	0.585	0.48	0.67	0.56	0.635	0.645	0.65	**0.79**	0.627
低相似度优先	0.69	0.65	0.7	0.59	0.72	0.69	0.6	0.64	0.660
随机选择	0.566	0.583	0.618	0.571	0.678	0.664	0.632	0.655	0.621
标签多样性	0.73	0.76	0.68	**0.72**	0.73	0.725	0.73	0.71	0.723
GED	0.540	0.565	0.7	0.63	0.72	0.61	0.64	0.67	0.634
实测验证	**0.785**	0.76	**0.72**	**0.72**	**0.77**	**0.765**	**0.78**	**0.79**	**0.761**
本文方案	**0.785**	**0.765**	0.7	**0.72**	**0.77**	**0.765**	0.75	0.71	0.746

联性能表现（其中**加粗**为最优，下划线为次优）。总体来看，本文提出的OSC-MM方法在大多数场景下取得了逼近实测验证策略的性能，同时整体优于仅依赖单一维度信息（能力或相似度）的匹配策略。

在CEVAL、MMLU和CMMLU等知识类与多选任务上，本文方法表现稳定。对于用户模型mi03，本文方法在CEVAL和MMLU上达到与最优组合一致的性能，在CMMLU上也接近最优水平；对于用户模型g405，本文方法在多个数据集上同样逼近或达到最优。这一结果表明，仅依靠能力标签互补进行筛选存在性能上限，而本文方法无需逐一验证候选模型，即可筛选出更高质量的协作模型。

相比之下，各基线策略均存在明显短板。能力优先策略虽整体表现较强，但在g405多相似模型场景下明显劣于最优组合，说明仅依赖模型能力难以捕捉协作所需的互补性；高相似度优先策略在多数任务中表现较差，尤其在用户模型g405场景中下降显著，验证了高同质性组合容易引入信息冗余。GED和低相似度优先策略虽在个别数据集上取得一定效果，但整体表现不足。其中，GED盲目追求极致的差异性（低Q值），说明单纯追求差异性并不能保证协作收益。标签多样性基线侧重于扩充能力标签覆盖范围，在g405场景下有小幅优势，但在mi03场景下性能表现不佳，可见仅靠标签维度筛选无法兼顾不同候选池分布。

针对两类用户模型场景的对比进一步验证了上述判断。在高同质性场景下，此时高相似度策略因信息冗余而导致性能显著下降；而在低同质性场景下，低相似度策略亦未能稳定提升性能。这表明，固定采用高相似或低相似的单一策略，均难以适应不同的模型分布条件。相比之下，本文方法在两类场景下均保持稳定表现，显示出根据用户模型特性自适应选择协作对象的能力。

在GSM8K数据集上，呈现出与其他任务不同的趋势：高相似度策略在部分情况下取得较优性能，而低相似度策略表现反而下降。这一现象主要原因是：在数学推理任务中，模型间的生成一致性对Token级并联具有重要影响——若协作模型差异过大，其推理路径和中间表达可能存在较大偏差，在Token级聚合过程中引入不稳定性，反而损害最终性能。值得注意的是，尽管单纯的高相似度策略在该任务上具有一定优势，但其在其他数据集上表现并不稳定。这说明，仅依赖相似度这一单一指标难以适应多样化的任务需求。

综合以上分析，并联模型匹配的关键在于在模型间的互补性与相似性之间取得平衡，仅依靠能力、相似度、错误和标签多样性任何单一维度，都存在显著的场景与任务短板；本文方案能实现更均衡、更贴近实测验证的并联性能。

3.6　效率分析

本节从验证规模和综合收益与计算开销权衡的角度，对本文提出的OSC-MM方法与基于逐一验证的实测方法进行对比分析。

实测验证方法需要将用户模型与候选池中所有模型逐一进行在线并联测试，其验证规模随模型总数N线性增长。同时，每次并联验证均要求两个模

型同时在线，在大规模场景下将带来较高的资源调度压力与系统负担。

实测表明，若将单次在线并联验证综合开销记为C_{on}。就计算任务本身而言，单次相似度计算的开销约为单次验证的0.6–0.7倍。同时，相比于双模型参与的在线并联验证，离线相似度计算仅依赖用户单模型，资源占用进一步折半。因此，单次离线相似度计算综合开销为$C_{off} = (0.3 \sim 0.35) \times C_{on}$。

3.6.1 验证规模

从验证规模的角度分析，假设模型集合中约有20%至50%的模型能够形成相对独立的簇结构（即$L \approx 0.2N \sim 0.5N$），综合考虑本文方法与实测验证方法在验证规模上的差异，分析结果如下：

$$\begin{cases} Cost_{实测} = N \times C_{on} \\ Cost_{本文} = L \times C_{off} + (L - 1) \times C_{on} \end{cases} \tag{10}$$

本文方法将匹配过程分解为两个阶段：离线处理与在线匹配。在离线阶段，预先计算模型间相似度并完成分簇，构建层次化的模型网络。当用户模型接入时，首先识别其归属簇并予以排除，避免选择输出行为高度相似的模型进行协作；随后，仅需与其余各独立簇中选取的Top模型进行在线并联实测。这一机制将匹配过程由“全量在线验证”转化为“可复用的离线建模+少量跨簇代表性验证”，验证规模由N降低至约L。

表6给出了不同模型规模N下对应的独立簇数量L的设置，该设置基于模型间相似性分布的合理假设：随着模型规模增加，新增模型更可能与已有模型存在相似性而被归入现有簇中，因此簇数量L的增长速度低于N，这符合实际大规模模型网络的结构特征。在该设定下，本文方法的在线并联验证次数降低为逐一验证方法的20%–50%，而验证开销的降低比例可达42.5%–68.0%，且这一优势随着模型规模的增大而更加明显。

3.6.2 综合收益与计算开销权衡

从综合收益与计算开销权衡的角度分析，本文构建了收益与开销的基准评估框架。为便于直观比较，均以“实测验证”方法为基准（即将其在测试集上取得的综合收益与计算开销均归一化为100%）。表7呈现了各匹配策略在不同候选池环境下的归一化表现。

在此框架中，综合收益定义为各方法在四个数据集上取得的平均准确率与实测基准的比值；综合开销则涵盖了从初期静态构建到后续动态选择协作模型的全过程模型资源开销与实测基准的比值。

在静态构建方面，“性能优先”等方法需测试所有模型在各验证集上的表现，因模型性能高度依赖数据集的领域与范围，故需跨多个验证集进行大规模测试。相反，基于相似度评估的方法旨在衡量模型间的输出相似性，其对验证集领域范围的依赖程度较低。

在动态选择方面，相似度计算的实际资源开销会受分簇数量、自有模型与簇内模型相似度分布状况的影响，因此其综合开销会呈现出一定的波动性。相比之下，实测验证等方法的计算部分较为固定，计算开销稳定。

由表7可见，本文方法在mi03、g405场景中均取得除实测验证外的最高任务收益，分别达到98.4%与97.5%。本文方法相比其余基线，无需大幅牺牲收益即可显著降低开销，且在两类环境下均具备稳定优异的收益和开销平衡。

综上，本文方法所带来的性能增益，并非以“额外增加开销”为代价；相反，即便计入预处理成本，其综合计算开销仍具备显著优势。此外，离线计算部分具有良好的可复用性。模型间相似度矩

表6 实测验证与本文方法验证规模对比

模型数量	独立簇数量	本文方法 离线相似度计算次数	本文方法在线并联验证次数	验证开销降低比例
10	5	5	4	42.5%-45.0%
20	8	8	7	51.0%-53.0%
50	15	15	14	61.5%-63.0%
100	30	30	29	60.5%-62.0%
150	40	40	39	64.7%-66.0%
200	50	50	49	66.8%-68.0%

表7　各方法在不同场景下的综合收益与计算开销对比

方法	综合收益		计算开销
	mi03	g405	
性能优先	96.6%	91.8%	32%
高相似度优先	83.1%	81.5%	24%-52%
低相似度优先	88.7%	84.7%	24%-52%
标签多样性	93.9%	96.0%	49%
GED	85.1%	81.5%	36%
本文方法	98.4%	97.5%	27%-68%
实测验证	100%	100%	100%

阵、分簇结构以及各模型的能力评估结果，均可作为模型网络的基础信息持续维护。在后续使用中，已接入的用户模型可作为网络中的固定节点参与调度，无需重复执行相似度计算与分簇过程，从而进一步降低长期使用成本。

4　结束语

多模型并联协作在模型互联网中具有重要应用价值。在协作模型匹配过程中，面临全量搜索带来的高昂计算开销，制约了其在大规模场景下的实际部署。为应对上述挑战，本文提出了一种基于相似度分簇的模型匹配方法OSC-MM。在保证协作效果的前提下，显著降低了相似度计算与并联验证的整体开销。实验结果表明，OSC-MM方法在多个数据集上取得了逼近实测验证的性能。此外，该方法在不同用户模型条件下均展现出良好的适应性与稳健性。本研究为模型互联网中大规模、高效率的协作模型匹配提供了可行路径。未来工作将重点研究模型网络的增量聚类与动态匹配策略，以应对模型动态变化对相似度结构的冲击，进一步提升OSC-MM方法在真实动态环境下的实用性与鲁棒性。

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(5): 1305-1318.
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(5): 1305-1318.

[2] 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.

[3] 王建辉, 李哲涛, 石伟凡, 等. 模型互联网中基于自我效能的Token级多模型协作[J]. 通信学报, 2026, 27(2): 125-139.
Wang J H, Li Z T, Shi W F, et al. Token-level multi-model collaboration based on self-efficacy in AI-model network[J]. Journal on Communications, 2026, 27(2): 125-139.

[4] 刘忠仁, 李哲涛, 王建辉, 等. 模型互联中多模型串并联协作推理[J]. 电子学报, 2025, 53(11): 3817-3835.
Liu Z R, Li Z T, Wang J H, et al. Multi-Model Serial and Parallel Collaborative Inference in AI-ModelNet[J]. Acta Electronica Sinica, 2025, 53(11): 3817-3835.

[5] 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 [EB/OL]. arXiv, 2024, arXiv: 2407.06089. https://arxiv. org/abs/2407.06089.

[6] Fu J, Jiang Y, Wu P, et al. Rethinking llm ensembling from the perspective of mixture models[EB/OL]. arXiv, 2026, arXiv:2605.00419. https://arxiv.org/abs/2605.00419.

[7] Cohen S, Cohen I N, Goldshlager N, et al. DFPE: A Diverse Fingerprint Ensemble for Enhancing LLM Performance[C]//Findings of the Association for Computational Linguistics: EACL 2026. Rabat, Morocco, March 24-29, 2026.

[8] Sourati Z, Ziabari A S, Dehghani M. The homogenizing effect of large language models on human expression and thought[EB/OL]. arXiv, 2025, arXiv:2508.01491. https://arxiv.org/abs/2508.01491v2.

[9] He Y, Wang T, Saad-Falcon A, et al. Supposedly equivalent facts that aren't? entity frequency in pre-training induces asymmetry in LLMs[C]//Proceedings of the Conference on Language Modeling (COLM 2025. Montreal, Canada, October 7-10, 2025.

[10] Kim E, Garg A, Peng K, et al. Correlated errors in large language models[C]//Proceedings of the 42nd International Conference on Machine Learning (ICML 2025. Vancouver Convention Center, Canada, July 13-19, 2025.

[11] Jiang L W, Chai Y J, Li M, et al. Artificial hivemind: The open-ended homogeneity of language models (and beyond)[C]//Advances in Neural Information Processing Systems (NeurIPS 2025. Santiago, Chile, December 2-7, 2025.

[12] Karouzos C, Tan X, Aletras N, et al. Where does output diversity collapse in post-training?[EB/OL]. arXiv, 2026, arXiv:2604.16027. https://arxiv.org/abs/2604.16027.

[13] Yang C, Holtzman A. LLM probability concentration: How alignment shrinks the generative horizon[C]//Proceedings of the International Conference on Machine Learning (ICML 2025. Vancouver Convention Center, Canada, July 13-19, 2025.

[14] Zhang T Y, Kishore V, Wu F, et al. BERTScore: Evaluating Text Generation with BERT[C]//Proceedings of the International Conference on Learning Representations (ICLR 2020. AbabaAddis, Ethiopia, 2020.

[15] Herrmann V, Alcaide E, Wand M, et al. Multiple Token Divergence: Measuring and Steering In-Context Computation Densi-ty[C]//Proceedings of the International Conference on Learning Representations (ICLR 2026).

[16] Xu Y F, 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).

Mexico City, Mexico, June 2024.
[17] Yao Y X, Wu H, Liu M Y, et al. Determine-Then-Ensemble: Necessity of Top-k Union for Large Language Model Ensembling[C]//The Thirteenth International Conference on Learning Representations (ICLR 2025). Singapore, April 24-28, 2025.
[18] Varangot-Reille C, Bouvard C, Gourru A, et al. Doing More with Less: A Survey on Routing Strategies for Resource Optimisation in Large Language Model-Based Systems[EB/OL]. arXiv, 2025, arXiv: 2502.00409. https://arxiv.org/abs/2502.00409.
[19] Shashidhar S, Chinta A, Sahai V, et al. Democratizing LLMs: An Exploration of Cost-Performance Trade-offs in Self-Refined Open-Source Models[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 9070-9084.
[20] Turkmen Y, Buyukates B, Bastopcu M. Don't Always Pick the Highest-Performing Model: An Information Theoretic View of LLM Ensemble Selection[EB/OL]. arXiv, 2026, arXiv: 2602.08003. https://arxiv.org/abs/2602.08003.
[21] Ong I, Almahairi A, Wu V, et al. RouteLLM: Learning to Route LLMs from Preference Data[C]//Proceedings of the International Conference on Learning Representations (ICLR 2025). Singapore, April 24-28, 2025. ICLR, 2025.
[22] Feng T, Shen Y Z, You J X. GraphRouter: A Graph-based Router for LLM Selections[EB/OL]. arXiv, 2024, arXiv:2410.03834. https://arxiv.org/abs/2410.03834.
[23] Cai R, Chen Z J, Wang Y F, et al. ModelLens: Finding the Best for Your Task from Myriads of Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025.
[24] Cruz R M O, Sabourin R, Cavalcanti G D C. Dynamic classifier selection: Recent advances and perspectives[J]. Information Fusion, 2018, 41: 195-216.
[25] Huang Y Z, Bai Y Z, Zhu Z H, et al. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models[C]//Advances in Neural Information Processing Systems (NeurIPS 2023. OrleansNew, LA, USA, 2023.
[26] Li H N, Zhang Y X, Koto F, et al. CMMLU: Measuring Massive Multitask Language Understanding in Chinese[EB/OL]. arXiv, 2023, arXiv: 2306.09212. https://arxiv.org/abs/2306.09212.
[27] Hendrycks D, Burns C, Basart S, et al. Measuring Massive Multitask Language Understanding[C]//Proceedings of the International Conference on Learning Representations (ICLR 2021. Vienna, Austria, 2021.
[28] Cobbe K, Kosaraju V, Bavarian M, et al. Training Verifiers to Solve Math Word Problems[C]//Advances in Neural Information Processing Systems (NeurIPS 2021). Virtual Conference, December 6-14, 2021.
[29] Yu Y J, Dai Z Y, Wang Z K, et al. OpenCSG Chinese Corpus: A Series of High-quality Chinese Datasets for LLM Training[EB/OL]. arXiv, 2025, arXiv:2501.08197. https://arxiv.org/abs/2501.08197.

龙赛琴（1986− ），女，博士，暨南大学信息科学技术学院教授，主要研究方向为模型互联网与算力网络。

赖俊杰（2001− ），男，暨南大学硕士生，主要研究方向为大模型与自然语言处理。

肖勇（2000− ），男，暨南大学博士生，主要研究方向为人工智能与数据隐私安全。

刘忠仁（1996− ），男，暨南大学博士生，主要研究方向为人工智能与大语言模型推理系统。

曾丽（1990− ），女，博士，长沙理工大学计算机科学与技术学院讲师，主要研究方向为可信人工智能、大语言模型与智能体。

李哲涛（1980− ），男，博士，暨南大学信息科学技术学院教授、博士生导师，主要研究方向为计算机网络、人工智能和安全。