

The Avengers: A Simple Recipe for Uniting Smaller Language Models to Challenge Proprietary Giants

Yiqun Zhang^{1,2♣‡} Hao Li^{2,3♣‡} Chenxu Wang^{2,4‡} Linyao Chen^{2,5‡}
 Qiaosheng Zhang² Peng Ye² Shi Feng^{1†} Daling Wang¹ Zhen Wang³
 Xinrun Wang⁶ Jia Xu² Lei Bai² Wanli Ouyang² Shuyue Hu^{2‡}
¹ Northeastern University ² Shanghai Artificial Intelligence Laboratory
³ Northwestern Polytechnical University ⁴ Beijing Institute of Technology
⁵ The University of Tokyo ⁶ Singapore Management University
 yiqunzhang@stumail.neu.edu.cn li.hao@mail.nwpu.edu.cn
 fengshi@cse.neu.edu.cn hushuyue@pjlab.org.cn

Abstract

Proprietary giants are increasingly dominating the race for ever-larger language models. Can open-source, smaller models remain competitive across a broad range of tasks? In this paper, we present the *Avengers*—a simple recipe that leverages the collective intelligence of these smaller models. The *Avengers* builds upon four lightweight operations: (i) *embedding*: encode queries using a text embedding model; (ii) *clustering*: group queries based on their semantic similarity; (iii) *scoring*: scores each model’s performance within each cluster; and (iv) *voting*: improve outputs via repeated sampling and voting. At inference time, each query is embedded and assigned to its nearest cluster. The top-performing model(s) within that cluster are selected to generate the response with repeated sampling. Remarkably, **with 10 open-source models (~7B parameters each), the *Avengers* surpasses GPT-4o, 4.1, and 4.5 in average performance across 15 diverse datasets spanning mathematics, coding, logical reasoning, general knowledge, and affective tasks.** In particular, it surpasses GPT-4.1 on mathematics tasks by 18.21% and on code tasks by 7.46%. Furthermore, the *Avengers* delivers superior out-of-distribution generalization, and remains robust across various embedding models, clustering algorithms, ensemble strategies, and values of its sole parameter—the number of clusters. We have open-sourced the code on GitHub: <https://github.com/ZhangYiqun018/Avengers>

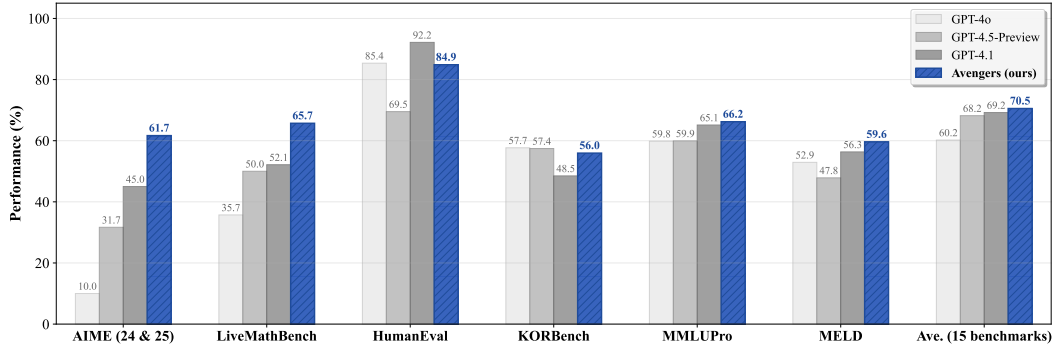


Figure 1: Comparison of proprietary models with our method (the *Avengers*) across six representative benchmarks. Full results on all 15 benchmarks are reported in Table 2.

[†] Corresponding author.

[♣] Equal contribution.

[‡] Work done during the author’s internship at Shanghai Artificial Intelligence Laboratory.

《复仇者联盟》：联合较小语言模型挑战专有巨头的简单配方

— 坤张浩李晨旭王林瑶陈启生张鹏叶世峰王达令王振新润王佳旭雷白万立欧阳舒月胡

¹ 东北大学 上海人工智能实验室
³ 西北工业大学 北京理工大学
⁵ 东京大学 新加坡管理大学
yiqunzhang@stumail.neu.edu.cn li.hao@mail.nwpu.edu.cn
fengshi@cse.neu.edu.cn hushuyue@pjlabor.org.cn

摘要

专有巨头在日益庞大的语言模型竞赛中占据主导地位。开源的小型模型能否在广泛任务中保持竞争力？在本文中，我们提出了 Avengers——一种利用这些小型模型集体智慧的简单方法。Avengers 基于四个轻量级操作：i) 嵌入：使用文本嵌入模型对查询进行编码；ii) 聚类：根据语义相似性对查询进行分组；iii) 评分：评估每个模型在每个聚类中的表现；iv) 投票：通过重复采样和投票改进输出。在推理时，每个查询被嵌入并分配到最近的聚类。该聚类中表现最佳的模型被选中，通过重复采样生成响应。

值得注意的是，仅使用 10 个开源模型（每个约 70 亿参数），Avengers 在涵盖数学、编程、逻辑推理、通用知识和情感任务的 15 个多样化数据集上的平均表现便超越了 GPT-4o、4.1 和 4.5。具体而言，它在数学任务上比 GPT-4.1 高出 18.21%，在编程任务上高出 7.46%。此外，Avengers 展现出更优的分布外泛化能力，并且在不同嵌入模型、聚类算法、集成策略及其唯一参数——聚类数量——的各种取值下均保持稳健。我们已在 GitHub 上开源代码：<https://github.com/ZhangYiqun018/Avengers>

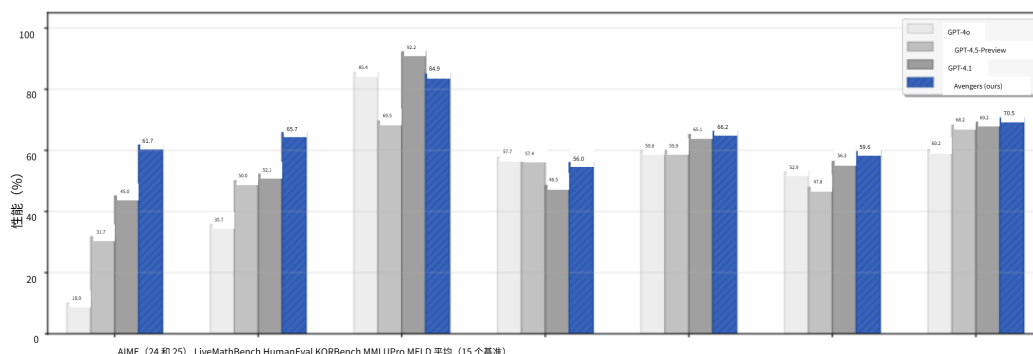


图 1：在六个代表性基准测试中，专有模型与我们的方法（复仇者联盟）的对比。所有 15 个基准测试的完整结果见表 2。

[†] 通讯作者。
[✳] 同等贡献。
[‡] 作者在上海人工智能实验室实习期间完成的工作。

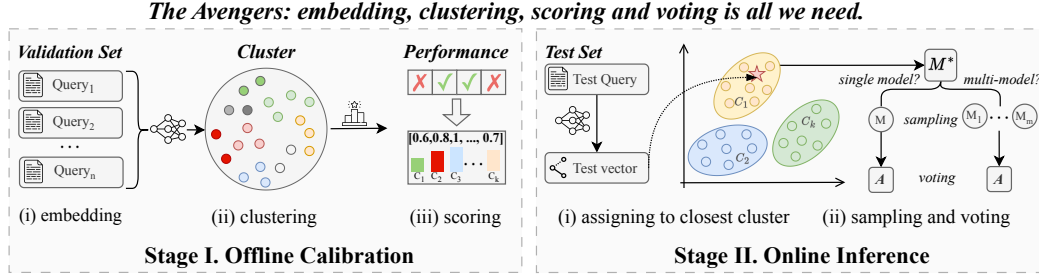


Figure 2: The Avengers, a simple recipe for *collective intelligence* of smaller language models.

1 Introduction

The current landscape of language models (LMs) is dominated by few proprietary giants. The race to build ever-larger models is far from over; however, it is becoming a closed competition, accessible only to organizations with massive resources. This trend risks marginalizing the broader research community and concentrating innovation within a privileged few. In this context, the open-source community—typically working with *smaller* LMs—must ask whether it is able to remain competitive. Although recent studies have shown that individual smaller models can excel in specific domains [1–4], the more pressing challenge is: can it remain competitive not only on narrowly defined tasks, but also in achieving comparable or even superior *overall* performance across a *broad* range of tasks?

A positive answer would have important implications. First, since open-source, smaller models are easily accessible and significantly more amenable, a positive answer would represent a meaningful step toward democratizing AI development [5]—enabling a broader community of researchers and practitioners to engage in and contribute to cutting-edge LM research. Second, the open-source ecosystem has already invested significant effort and computational resources in model development.¹ Demonstrating strong performance from these models would justify the reuse of existing efforts, supporting more sustainable AI research [6]. Third, a positive answer would pave a new path for advancing LM research, providing another viable alternative to the ever increasing of model size.

We posit that the key to addressing the challenge lies in *collective intelligence*—while a single smaller model may not match the generality of larger ones, the collective performance of a team of multiple smaller models may prove competitive. Nevertheless, operationalizing collective intelligence is non-trivial and involves several important design considerations. Since it is infeasible to incorporate all available models in the community, which subset of models should be selected? How should we orchestrate this subset of models to maximize their collective performance? For a notable example, for each incoming query, should we rely on a router-based approach [7–10] that selects the most capable model, or should we adopt a mixture-based strategy (e.g., Mixture-of-Agents [11–14]) where multiple models respond in parallel and their outputs are aggregated into a final answer? Moreover, given the rapid evolution of the LM landscape, with new tasks and new models continuously emerging, how can such systems adapt to these ongoing changes?

To this end, we introduce the *Avengers*—a simple recipe that effectively leverages the collective intelligence of smaller LMs. Given a set of models and a dataset (potentially spanning multiple tasks and benchmarks), the *Avengers* operates as follows. First, it analyzes the tasks by embedding queries from the validation set and clustering them based on their semantic representations. Next, to identify each model’s strengths and weaknesses, it evaluates each model on the validation set and constructs a cluster-wise capability profile—a simple, numeric vector representing model performance across clusters. At inference time, for each query in the test set, the *Avengers* computes its embedding and assigns it to the nearest cluster. Based on the capability profiles, the model(s) with the highest performance in that cluster are then selected to handle the query. The final output is generated using a repeated sampling and voting strategy over the selected model(s). An overview of the *Avengers* is presented in Figure 2.

¹For example, as of May 2025, there are over 1,318 models fine-tuned on LLaMA-3.1-8B-Instruct and 489 on Qwen-2.5-7B-Instruct, despite both being released less than a year ago.

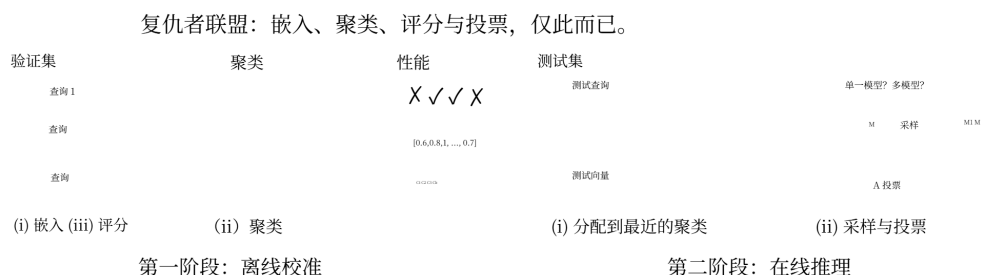


图 2：复仇者联盟，一种实现较小语言模型集体智能的简单方法。

1 引言



当前语言模型（LM）的格局由少数专有巨头主导。构建更大模型的竞赛远未结束；然而，它正变成一场封闭的竞争，只有拥有庞大资源的组织才能参与其中。这一趋势有可能使更广泛的研究社区被边缘化，并将创新集中在少数特权者手中。在此背景下，开源社区——通常使用较小的 LM——必须自问是否能够保持竞争力。尽管近期研究表明，单个较小模型可以在特定领域表现出色[1–4]，但更紧迫的挑战是：它能否不仅在狭义定义的任务上保持竞争力，还能在广泛任务范围内实现可比甚至更优的整体性能？

肯定的答案将具有重要影响。首先，由于开源的小型模型易于获取且适应性更强，肯定的答案将代表着向 AI 开发民主化迈出的有意义一步[5]——使更广泛的研究人员和实践者群体能够参与并贡献于前沿语言模型研究。其次，开源生态系统已在模型开发上投入了大量精力和计算资源。证明这些模型具有强大性能将证明复用现有努力的合理性，支持更可持续的 AI 研究[6]。第三，肯定的答案将为推进语言模型研究开辟新路径，为不断增大的模型规模提供另一种可行的替代方案。

我们认为，应对这一挑战的关键在于集体智慧——尽管单个较小模型可能无法与较大模型的通用性相匹敌，但多个较小模型组成的团队的集体表现或许能够具备竞争力。然而，将集体智慧付诸实践并非易事，涉及若干重要的设计考量。由于将社区中所有可用模型全部纳入并不可行，那么应选择哪些模型的子集？应如何编排这一模型子集以最大化其集体表现？举一个显著的例子：对于每个传入的查询，我们是应依赖基于路由的方法[7–10]来选择最有能力的模型，还是应采用基于混合的策略（例如，Mixture-of-Agents [11–14]），让多个模型并行响应并将其输出聚合为最终答案？此外，鉴于语言模型领域快速演进，新任务和新模型不断涌现，此类系统应如何适应这些持续变化？

为此，我们引入了 Avengers——一种简单的方法，有效利用较小语言模型的集体智慧。给定一组模型和一个数据集（可能涵盖多个任务和基准测试），Avengers 的操作如下。首先，它通过嵌入验证集中的查询并根据其语义表示进行聚类来分析任务。接下来，为了识别每个模型的优势与不足，它在验证集上评估每个模型，并构建一个按聚类划分的能力概况——一个简单的数值向量，表示模型在各聚类中的表现。在推理时，对于测试集中的每个查询，Avengers 计算其嵌入并将其分配到最近的聚类。基于能力概况，选择在该聚类中表现最佳的模型来处理该查询。最终输出通过对所选模型进行重复采样和投票策略生成。Avengers 的概览如图 2 所示。

¹例如，截至 2025 年 5 月，基于 LLaMA-3.1-8B-Instruct 微调的模型超过 1,318 个，基于 Qwen-2.5-7B-Instruct 微调的模型有 489 个，尽管这两个模型发布至今均不到一年。

The *Avengers* is straightforward to implement, requiring no neural network training. It is highly reproducible, avoiding reliance on hand-crafted prompts or architecture-specific model choices. It is plug-and-play, allowing seamless integration of newly available models by incrementally evaluating them on the validation set and computing their capability profiles. Moreover, it is well-suited for cold-start, assuming no human expertise or prior knowledge for model set composition; given a large candidate model pool, the construction of capability profiles enables automatic selection of high-performing models while preserving diversity. Conceptually, the *Avengers* can be viewed as a routing-based approach. However, as detailed in the Related Work section and summarized in the qualitative comparison in Table 1, it differs significantly from prior work in this category—both methodologically, through its clustering-based routing strategy, and empirically, in the notably strong performance it achieves.

Our experiments show that, the *Avengers*, using 10 models ($\sim 7B$ each), surpasses the average performance of GPT-4.1² across 15 datasets spanning five domains (mathematics, code, logic, knowledge, and affective tasks). Notably, the *Avengers* outperforms GPT-4.1 on nine of these datasets; it outperforms GPT-4.1 by 37.04% on AIME, by 31.86% on MBPP, and by 15.41% on KORBench. When isolating the effect of the routing mechanism, the *Avengers*’ clustering-based router still outperforms state-of-the-art (SOTA) performance-oriented routers (RouterDC [8], embedLLM [9], and MODEL-SAT [10]), even though all those routers require neural network training (with some even involving fine-tuning LMs) and must be retrained to accommodate new models or tasks. On out-of-distribution tasks (evaluated on five additional benchmarks), the superior performance of the *Avengers*’ clustering-based router is even more noticeable; it outperforms RouterDC by 10.31%, embedLLM by 8.14%, and MODEL-SAT by 8.56%, showcasing its strong generalization capability.

Additional experiments demonstrate that the *Avengers* is compatible with a variety of embedding models, clustering algorithms, and ensemble strategies for aggregating outputs from selected models. It maintains robust performance across different clustering methods (K-Means [15, 16], Hierarchical Clustering [17], GMM [18], Spectral Clustering [19] and BIRCH [20]). Furthermore, its performance can be further enhanced by employing stronger embedding models and ensemble strategies that explicitly leverage multiple models, such as the multi-model variant [21] of the repeated-sampling-and-voting approach. Furthermore, we show that the *Avengers* can automatically construct an effective model set given a predefined set size. Given a candidate pool of 22 models, it can match the performance of GPT-4.1 with as few as three selected models, yet its performance continues to improve as the number of models increases, particularly in Code, Knowledge, and Affective tasks (Figure 3). In addition, regarding data efficiency, we show that the *Avengers* continues to outperform GPT-4.1 even when only 30% of the data is used for validation, leaving the remaining 70% for testing (Figure 8). Finally, we show that the *Avengers* is robust to variations in its sole hyperparameter, i.e. the number of clusters k . Building on this observation, we offer a hypothesis to explain why clustering-based routing can outperform typical supervised routing methods, despite requiring no training (Appendix B.4, Figure 4).

An exploratory analysis (Appendix B.5, Figure 9)) about the working mechanism of the *Avengers* reveals that it can automatically select models that strike a favorable balance between strong individual performance and diversity, which lays the foundation for accurate query-to-model routing. From a candidate pool, it can select models, such as Fin-R1 and LLaMA-3.1-8B-UltraMedical, which are of rather low visibility in the community. However, these models prove crucial to overall performance. Notably, $\sim 50\%$ of queries from the ARC Challenge are routed to fin-R1 (originally developed for financial tasks), and $\sim 40\%$ of queries from MedQA are routed to LLaMA-3.1-8B-UltraMedical (developed by a small group of university students). These results not only demonstrate the effectiveness of the *Avengers*, but also more importantly, highlight the untapped potential of open-source contributions—advancing the vision of a more democratic and inclusive AI ecosystem.

2 Related Work

The harnessing of collective intelligence from multiple models constitutes one of the frontiers of AI and ML research, and has recently attracted much interest [22–27]. Existing approaches in this area generally fall into three paradigms: router-based, mixture-based, and merging-based methods. This

²GPT-4.1: OpenAI’s flagship model released on April 14, 2025.

Avengers 实现简单直接，无需进行神经网络训练。它具有高度可复现性，避免依赖手工设计的提示词或特定架构的模型选择。它即插即用，能够在验证集上增量评估新发布的模型并计算其能力画像，实现无缝集成。此外，它非常适合冷启动场景，无需人工专业知识或先验知识来组合模型集合；在给定大型候选模型池的情况下，能力画像的构建能够自动选择高性能模型，同时保持多样性。从概念上讲，Avengers 可被视为一种基于路由的方法。然而，如相关工作部分所述及表 1 中的定性比较所总结，它与该类别中的先前工作存在显著差异——无论是在方法论上（通过其基于聚类的路由策略），还是在实证上（其表现出的显著强劲性能）。

我们的实验表明，Avengers 使用 10 个模型（每个约 7B 参数）在涵盖数学、代码、逻辑、知识和情感任务五个领域的 15 个数据集上超越了 GPT-4.1 的平均性能。值得注意的是，Avengers 在其中 9 个数据集上优于 GPT-4.1；在 AIME 上超越 GPT-4.1 达 37.04%，在 MBPP 上达 31.86%，在 KORBench 上达 15.41%。当单独评估路由机制的效果时，Avengers 基于聚类的路由器仍然优于最先进的性能导向路由器（RouterDC [8]、embedLLM [9] 和 MODEL-SAT [10]），尽管所有这些路由器都需要神经网络训练（有些甚至涉及对语言模型的微调），并且必须重新训练以适应新模型或任务。在分布外任务上（在五个额外基准上评估），Avengers 基于聚类的路由器的优越性能更加明显；它比 RouterDC 高出 10.31%，比 embedLLM 高出 8.14%，比 MODEL-SAT 高出 8.56%，展示了其强大的泛化能力。

额外的实验表明，Avengers 与多种嵌入模型、聚类算法以及用于聚合所选模型输出的集成策略兼容。它在不同的聚类方法（K-Means [15, 16]、层次聚类 [17]、GMM [18]、谱聚类 [19] 和 BIRCH [20]）下均保持稳健的性能。此外，通过采用更强的嵌入模型和显式利用多个模型的集成策略（例如重复采样与投票方法的多模型变体 [21]），其性能可以进一步提升。我们还表明，Avengers 能够在给定预定义集合大小的情况下自动构建有效的模型集合。在包含 22 个模型的候选池中，它仅需选择三个模型即可匹配 GPT-4.1 的性能，但随着模型数量的增加，其性能持续提升，尤其是在代码、知识和情感任务中（图 3）。此外，在数据效率方面，我们表明即使仅使用 30% 的数据进行验证，剩余 70% 用于测试，Avengers 仍能超越 GPT-4.1（图 8）。最后，我们展示了 Avengers 对其唯一超参数，即聚类数量 k 的变化具有鲁棒性。基于这一观察，我们提出了一种假设，用以解释为何基于聚类的路由方法能够超越典型的监督式路由方法，尽管它无需训练（附录 B.4，图 4）。

关于复仇者联盟工作机制的探索性分析（附录 B.5，图 9）表明，它能够自动选择在强大个体性能与多样性之间取得良好平衡的模型，这为精确的查询到模型路由奠定了基础。从候选池中，它能够选择诸如 Fin-R1 和 LLaMA-3.1-8B-UltraMedical 等社区知名度较低的模型。然而，这些模型对整体性能至关重要。值得注意的是，ARC Challenge 中约 50% 的查询被路由至 fin-R1（最初为金融任务开发），而 MedQA 中约 40% 的查询被路由至 LLaMA-3.1-8B-UltraMedical（由一小群大学生开发）。这些结果不仅证明了复仇者联盟的有效性，更重要的是，凸显了开源贡献尚未开发的潜力——推进了构建更加民主和包容的 AI 生态系统的愿景。

2 相关工作

多模型集体智能的利用构成了 AI 与 ML 研究的前沿领域之一，近期引起了广泛关注[22–27]。该领域的现有方法通常分为三种范式：基于路由、基于混合和基于合并的方法。这

² GPT-4.1: OpenAI 于 2025 年 4 月 14 日发布的旗舰模型。

work is most closely aligned with the router-based paradigm. Due to space constraints, we focus our discussion on this line of research and defer coverage of the other two paradigms to the Appendix.

The main goal of most router-based methods is to enhance the overall performance of a set of smaller models through query routing; a neural network-based router is often trained to select, for each incoming query, the model most capable of handling it [28–33]. LLM-Blender [7] utilizes pairwise comparisons to select the top- k models for each query and then fuses their outputs to enhance performance. ZOOTO [34] proposes reward-guided query routing, employing tag-based label enhancement to improve training stability. More recently, RouterDC [8], which employs dual contrastive learning to enhance routing accuracy, and EmbedLLM [9] leverages learned compact model embeddings along with query embeddings to predict routing correctness. Additionally, Model-SAT [10] generates capability instructions from model aptitude outcomes and employs text-aligned embeddings to guide a lightweight LLM in selecting optimal candidate models.

In parallel with the above performance-oriented research, a few recent studies have also explored routing strategies that account for the trade-off between performance and computational cost. Routellm [30] trains a binary classifier using preference data to dynamically route queries to either a stronger or a weaker LLM during inference. GraphRouter [31] constructs a heterogeneous graph comprising task, query, and LLM nodes, and predicts the performance-cost score via an edge prediction mechanism. RouterBench [35] introduces a suite of benchmarks along with several routing strategies. One such strategy routes each query to an LLM based on the average cost-penalized performance of the query’s k nearest neighbors. Similarly, [36] introduces a cluster-based approach that groups queries using a large unlabeled dataset with K -means; each query is assigned to a cluster, and an LLM is selected based on cost-adjusted per-cluster error scores computed from another labeled dataset.

The *Avengers* falls within the router-based paradigm but departs significantly from prior work in two important ways. First, unlike most existing methods that rely on training neural routers, which must be re-trained to accommodate new tasks or models—a significant limitation given the rapid pace of LLM development. In contrast, our approach is entirely *training-free*, yet remains highly adaptable and competitive, offering a lightweight, reproducible alternative without sacrificing performance. Second, while training-free routing (e.g., K -NN or K -means routing) has been explored in cost-performance tradeoff settings, they typically operate on model pools with large size disparities (often including models far larger than those used in our work), yet still *fail* to demonstrate competitive performance. To the best of our knowledge, this is the first work to show that a training-free, router-based framework—when integrated with automatic complementary model selection and voting—can elevate small open-source models to match the *overall* performance of a proprietary, flagship LLM.

3 The Avengers

In this section, we introduce the *Avengers*, a simple yet effective framework for harnessing the collective intelligence of multiple smaller language models. The *Avengers* is based on the observation that the open-source ecosystem offers a wide range of small models, with growing evidence that these models exhibit substantial diversity. This diversity arises in two key ways: (i) some models are domain-specific experts, fine-tuned for particular tasks such as mathematical reasoning; and (ii) others, while general-purpose, differ due to variations in training data, architectures, and design choices made by different stakeholders.

This diversity provides a strong foundation for collective intelligence. However, operationalizing it requires careful consideration of several key design decisions. A core principle underlying the *Avengers* is the idea of “**horses for courses**”—selecting the most suitable model(s) for each incoming query and routing the query to that model. We prioritize a routing strategy over a mixture strategy,³ because unlike larger models, smaller models tend to be more sensitive to prompt variations and often struggle with instruction following. Prior work has shown that mixture-based methods, while effective with larger models, can lead to degraded performance when applied to smaller ones.

The question then becomes: how do we identify the right “horses” for the right “courses”? This requires understanding both (i) what models (“horses”) are good at, and (ii) what types of queries (“courses”) are being handled. Moreover, it also necessitates effective matching—routing each query

³Mixture-based methods often demonstrate greater performance gains over single models, compared to router-based approaches.

我们的工作与基于路由器的范式最为接近。由于篇幅限制，我们重点讨论这一研究方向，并将另外两种范式的覆盖内容推迟至附录中。

大多数基于路由的方法的主要目标是通过查询路由来提升一组较小模型的整体性能；通常会训练一个基于神经网络的路由器，为每个传入查询选择最能处理该查询的模型 [28–33]。LLM-Blender [7] 利用成对比较为每个查询选择前 k 个模型，然后融合其输出以提升性能。ZOOTER [34] 提出了奖励引导的查询路由，采用基于标签的标签增强来提高训练稳定性。最近，RouterDC [8] 采用双重对比学习来提升路由准确性，而 EmbedLLM [9] 利用学习到的紧凑模型嵌入以及查询嵌入来预测路由正确性。此外，ModelSAT [10] 从模型能力结果中生成能力指令，并采用文本对齐的嵌入来引导轻量级 LLM 选择最优候选模型。

在与上述以性能为导向的研究并行推进的同时，近期也有少量研究探索了兼顾性能与计算成本权衡的路由策略。Routellm [30] 利用偏好数据训练一个二元分类器，在推理过程中动态地将查询路由至较强或较弱的 LLM。GraphRouter [31] 构建了一个包含任务、查询和 LLM 节点的异构图，并通过边预测机制来预测性能-成本得分。RouterBench [35] 引入了一套基准测试以及多种路由策略，其中一种策略根据查询的 k 个最近邻的平均成本惩罚性能，将每个查询路由至某个 LLM。类似地，[36] 提出了一种基于聚类的方法，利用大型无标签数据集通过 K-means 对查询进行分组；每个查询被分配至一个聚类，并基于另一个有标签数据集计算出的按聚类划分的成本调整误差得分来选定 LLM。

复仇者联盟属于基于路由器的范式，但在两个重要方面与先前工作有显著不同。首先，与大多数依赖训练神经路由器（必须重新训练以适应新任务或模型，鉴于 LLM 发展的快速步伐，这是一个重大限制）的现有方法不同，我们的方法完全无需训练，却依然高度适应且具有竞争力，提供了一种轻量级、可复现的替代方案，且不牺牲性能。其次，虽然无需训练的路由（例如 K-NN 或 K-means 路由）已在成本-性能权衡设置中有所探索，但它们通常作用于规模差异较大的模型池（通常包含远大于我们工作中所用模型的模型），却仍未能展示出有竞争力的性能。据我们所知，这是首个证明无需训练、基于路由器的框架——当与自动互补模型选择和投票集成时——能够将小型开源模型提升至与专有旗舰 LLM 整体性能相匹配的工作。

3 复仇者联盟

在本节中，我们介绍 Avengers，一个简单而有效的框架，用于汇聚多个较小语言模型的集体智慧。Avengers 基于这样一个观察：开源生态系统提供了多种多样的小型模型，且越来越多的证据表明这些模型展现出显著的多样性。这种多样性以两种关键方式产生：(i) 一些模型是特定领域的专家，针对数学推理等特定任务进行了微调；(ii) 其他模型虽然通用，但由于不同利益相关者在训练数据、架构和设计选择上的差异而有所不同。

这种多样性为集体智能提供了坚实的基础。然而，要将其付诸实践，需要仔细考虑几个关键的设计决策。Avengers 的核心原则是“因地制宜”的理念——为每个传入的查询选择最合适的模型，并将查询路由到该模型。我们优先采用路由策略而非混合策略，因为与较大的模型不同，较小的模型往往对提示变化更为敏感，并且在遵循指令方面常常存在困难。先前的研究表明，基于混合的方法虽然对较大模型有效，但在应用于较小模型时可能导致性能下降。

接下来的问题是：我们如何为正确的“课程”识别出正确的“马匹”？这需要理解 (i) 哪些模型（马匹）擅长什么，以及 (ii) 正在处理哪些类型的查询（课程）。此外，这还需要有效的匹配——将每个查询路由

³基于混合的方法相较于基于路由的方法，通常在性能提升上优于单一模型。

to the model best suited for its type. Consider a set M of models and a set $\mathcal{D}$ of query-answer sample pairs, which can be split into a validation set $\mathcal{D}_{\text{val}}$ and a test set $\mathcal{D}_{\text{test}}$. The *Avengers* address this question in two stages: (i) **offline calibration**, where it builds an understanding of both query types and model capabilities based on $\mathcal{D}_{\text{val}}$, and (ii) **online inference**, where it dynamically routes each query from $\mathcal{D}_{\text{test}}$ to the model it deems most capable and enhances the final output.

Offline Calibration During this stage, the *Avengers* constructs a structured understanding of the query type via embedding and clustering, and characterizes model capabilities through cluster-wise performance scores. First, to characterize query types, each query $d \in \mathcal{D}_{\text{val}}$ is encoded using a text **embedding** model, producing a semantic vector representation. Then, these vectors are clustered into k groups with the use of a **clustering** method (e.g., K -means), which results in a set $\mathcal{C}$ of k distinct clusters. Each cluster $c \in \mathcal{C}$ represents a semantically coherent group of queries, i.e., a query type. Next, to characterize model capabilities, each model $m \in \mathcal{M}$ is evaluated on $\mathcal{D}_{\text{val}}$, and its performance score is recorded within each cluster. This yields a cluster-wise **capability profile** for each model, which can be represented as a vector $p = [p_1, \dots, p_k]$, where p_i denotes the model’s performance score on cluster $c_i \in \mathcal{C}$. These profiles inform routing decisions at the inference time.

Online Inference At the inference time, the *Avengers* no longer performs clustering or profiling. For each query in the test set $\mathcal{D}_{\text{test}}$, the *Avengers* first computes its **embedding** using the same text embedding model used during offline calibration. The query is then **routed** to the nearest cluster $c_* \in \mathcal{C}$, based on distance in the embedding space. Given the nearest cluster c_* , the *Avengers* consult the capability profiles and select the top- n model(s) with the highest performance score p_* within that cluster. Notably, the selected model(s) is not necessarily the one with the best overall performance on $\mathcal{D}_{\text{val}}$ or $\mathcal{D}_{\text{test}}$. Rather, it is selected for its strength in the specific query type (or cluster). Once the top- n model(s) is selected, the *Avengers* generates responses using **repeated sampling**, followed by majority **voting** to determine the final output. That is, when $n = 1$, it adopts the Self-Consistency (SC) [37]. When $n > 1$, it adopts the Model-Switch [21], a multi-model, sample-efficient extension of SC.

Automatic Model Set Construction It is worth mentioning that the composition of the model set $\mathcal{M}$ can significantly influence collective performance. Intuitively, strong overall performance is unlikely to emerge from a group of models that underperform across all tasks. While prior approaches often rely on human expertise or domain knowledge to curate such model sets, the *Avengers* automates the selection of models with complementary strengths. Given a predefined deployment budget of $|\mathcal{M}|$ models and a larger candidate pool $\mathcal{M}'$, the *Avengers* enables automatic model selection from $\mathcal{M}'$. During the offline calibration stage, for each candidate model $m' \in \mathcal{M}'$, the *Avengers* computes the cluster-wise capability profile and calculates an overall score across clusters: $s(m') = \sum_{c \in \mathcal{C}} 1/r_{m'}^c$, where $r_{m'}^c$ denotes the rank of model m' within cluster c , determined by its cluster-wise performance. A higher score $s(m')$ indicates that the model either excels in specific clusters (yielding high ranks) or maintains strong performance across multiple query types. By selecting the top $|\mathcal{M}|$ models with the highest overall scores, the *Avengers* naturally constructs a complementary model set $\mathcal{M}$ that balances specialization and diversity across different query types.

New Datasets and Newly Available Models The fast-paced evolution of LM research regularly introduces new datasets, tasks, and models. While prior routing-based methods often require retraining neural networks to accommodate such changes, the *Avengers* is designed to be highly adaptable. To incorporate a new, out-of-distribution dataset $\mathcal{D}'$, the *Avengers* re-executes the offline calibration stage. This requires only *incremental* re-evaluation of existing models $\mathcal{M}$ on the validation set $\mathcal{D}'_{\text{val}}$, followed by re-clustering the queries from $\mathcal{D}'_{\text{val}} \cup \mathcal{D}_{\text{val}}$ based on updated query embeddings. Incorporating a newly available model $m^\dagger$ is even more lightweight. The *Avengers* simply *incrementally* evaluates the new model $m^\dagger$ on the existing validation set $\mathcal{D}_{\text{val}}$ and then computes its cluster-level performance profile, even without re-clustering.

4 Experiments

4.1 Experimental Setup

Datasets We primarily consider 15 datasets covering five categories: **Mathematics** (AIME, Math500 [38], and LiveMathBench [39]), **Code** (MBPP [40] and HumanEval [41]), **Logic** (KORBench [42], Knights and Knaves [43], and BBH [44]), **Knowledge** (ARC Challenge [45],

给最适合其类型的模型。考虑一组模型 M 和一组查询-答案样本对 D ，可将其分为验证集 D 和测试集 D 。复仇者联盟分两个阶段解决这一问题：i) 离线校准，在此阶段基于 D 构建对查询类型和模型能力的理解；ii) 在线推理，在此阶段动态地将来自 D 的每个查询路由到其认为最有能力的模型，并增强最终输出。

离线校准 在此阶段，Avengers 通过嵌入和聚类构建对查询类型的结构化理解，并通过聚类级性能评分刻画模型能力。首先，为刻画查询类型，每条查询 $d \in D$ 使用文本嵌入模型进行编码，生成语义向量表示。随后，利用聚类方法（如 K-means）将这些向量聚为 k 组，得到包含 k 个不同聚类的集合 C 。每个聚类 $c \in C$ 代表一组语义连贯的查询，即一种查询类型。接下来，为刻画模型能力，每个模型 $m \in M$ 在 D 上进行评估，并记录其在每个聚类内的性能评分。由此得到每个模型的聚类级能力画像，可表示为向量 $p = [p_1, \dots, p_k]$ ，其中 p_i 表示模型在聚类 $c_i \in C$ 上的性能评分。这些画像在推理时用于指导路由决策。

在线推理阶段，复仇者联盟不再执行聚类或画像分析。对于测试集 D 中的每个查询，复仇者联盟首先使用与离线校准阶段相同的文本嵌入模型计算其嵌入表示。随后，根据嵌入空间中的距离，将该查询路由至最近的聚类 $c \in C$ 。在确定最近的聚类 c 后，复仇者联盟查阅能力画像，并选择该聚类内性能得分 p 最高的前 n 个模型。值得注意的是，所选模型不一定是整体性能在 D 或 D 上最优的模型，而是根据其在特定查询类型（或聚类）中的优势进行选择。一旦选定前 n 个模型，复仇者联盟通过重复采样生成响应，并采用多数投票机制确定最终输出。即当 $n=1$ 时，采用自一致性（SC）[37]方法；当 $n>1$ 时，采用模型切换（Model-Switch）[21]方法，这是 SC 的一种多模型、样本高效的扩展。

自动模型集构建 值得一提的是，模型集 M 的组成会显著影响整体性能。直观来看，一组在所有任务上表现不佳的模型很难产生强大的整体性能。虽然先前的方法通常依赖人类专家或领域知识来策划此类模型集，但 Avengers 自动化了具有互补优势的模型选择过程。给定预定义的部署预算 $|M|$ 个模型和一个更大的候选池 M ，Avengers 能够从 M 中自动选择模型。在离线校准阶段，对于每个候选模型 $m \in M$ ，Avengers 计算簇级能力画像，并跨簇计算总体得分： $s(m) =$

$$\frac{1}{|C|} \sum_{c \in C} \frac{1}{r_{m,c}}$$

其中 $r_{m,c}$ 表示模型 m 在簇 c 内的排名，由其簇内性能决定。较高的得分 $s(m)$ 表明该模型要么在特定簇中表现优异（获得较高排名），要么在多种查询类型中保持强劲性能。通过选择总得分最高的前 $|M|$ 个模型，Avengers 自然构建了一个互补的模型集合 M ，该集合在不同查询类型之间平衡了专业性与多样性。

新数据集与新可用模型 语言模型研究的快速演进不断引入新的数据集、任务和模型。虽然先前的基于路由的方法通常需要重新训练神经网络以适应此类变化，但 Avengers 被设计为具有高度适应性。为了纳入新的分布外数据集 D ，Avengers 会重新执行离线校准阶段。这仅需在验证集 D 上对现有模型 M 进行增量式重新评估，然后基于更新的查询嵌入对来自 $D \cup D$ 的查询重新聚类。纳入新可用模型 m 则更为轻量。Avengers 只需在现有验证集 D 上对模型 m 进行增量式评估，然后计算其聚类级性能概况，甚至无需重新聚类。

4 实验

4.1 实验设置

数据集 我们主要考虑 15 个数据集，涵盖五个类别：数学（AIME、Math500 [38] 和 LiveMathBench [39]）、代码（MBPP [40] 和 HumanEval [41]）、逻辑（KORBench [42]、Knights and Knaves [43] 和 BPL [44]）、知识（ARC Challenge [45]）

MMLUPro [46], GPQA [47], FinQA [48], and MedQA [49]), and **Affective** (EmoryNLP [50] and MELD [51]). Additionally, to evaluate the generalization capabilities of the *Avengers* on out-of-distribution (OOD) tasks, we further include one extra dataset for each category: MathBench [52], StudentEval [53], Winogrande [54], BRAINTEASER [55], and DailyDialog [56]. Note that the additional five datasets are used exclusively for evaluating the OOD generalization. Detailed information about these datasets is provided in Appendix A.3 and Table 9.

Implementation of the *Avengers* We uniformly employ *gte-qwen2-7B-instruct* as the embedding model and utilize the classic K -Means with the number of clusters set to $K = 64$. Each dataset is randomly split into 70% for fitting cluster centroids and the remaining 30% for performance evaluation. Initially, we consider a pool consisting of 22 open-source candidate LLMs ($\sim 7B$), from which $m = 10$ models are automatically selected. During inference, we adopt the Self-Consistency (SC) strategy by default, setting the number of sampling rounds to 10. All experiments are repeated five times using five random seeds (42, 999, 2024, 2025, and 3407), and average results are reported. Detailed model information and deployment specifics can be found in the Appendix A.2.

Baselines We select baseline methods from two distinct technical directions (mixture- and router-based) to comprehensively evaluate the effectiveness of the *Avengers*. Additionally, to provide intuitive performance benchmarks, we include three proprietary models, **gpt-4o-2024-08-06** (GPT-4o), **gpt-4.5-preview** (GPT-4.5-preview) and **gpt-4.1-2025-04-14** (GPT-4.1). Specifically, we consider the following baseline methods:

- **Random Router:** Randomly selects one model from candidate set S for each query.
- **LLM Router:** A prompt-based router that uses an LLM (Qwen2.5-7B-Instruct) to select models based on natural-language profiles summarizing candidate model performance.
- **RouterDC** [8]: Utilizes dual contrastive learning to train embeddings for queries and models. The contrastive objectives ensure query embeddings are close to suitable models and semantically similar queries cluster together.
- **EmbedLLM** [9]: An embedding-based router trained via binary cross-entropy loss to predict model-query compatibility.
- **MODEL-SAT** [10]: Employs capability instruction tuning, translating candidate models’ performance into textual capability descriptions, which are embedded and passed to a trainable LLM to predict model suitability for each query.
- **Mixture of Agents (MoA)** [11]: A hierarchical mixture approach where multiple layers of LLMs iteratively refine or aggregate candidate answers, culminating in a final aggregation step.
- **Symbolic-MoE (SMoE)** [14]: A mixture method using skill keyword matching. Candidate models are selected based on query-derived skill keywords, and their outputs are aggregated via a single-layer MoA.

The methodological differences between these baseline methods and the *Avengers* are summarized in Table 1. More detailed implementation specifics can be found in Appendix A.4.

Enhancement to the Baselines Notably, to better align the baseline methods with our experimental scenarios and facilitate comprehensive comparisons, we introduce targeted enhancements for certain baselines. All baseline methods utilize **the same set of 10 models** automatically selected by the *Avengers* framework. Specifically, for the Mixture-based method MoA, we additionally evaluate an oracle variant—MoA (Oracle)—in which we manually **select the top-3 performing models** per task instead of using all 10 models. This approach addresses the performance degradation arising from excessively long context windows in the original MoA setup. For the trainable router-based methods (RouterDC, EmbedLLM, and MODEL-SAT), we do not employ a fixed training step count as a stopping criterion; instead, we report each method’s performance at its **peak on the test set**. Additionally, all SLMs and router-based results reported in Table 2 and 3 **incorporate the same Self-Consistency (SC)** strategy used by the *Avengers*, with 10 samples per query, ensuring fairness and robustness across comparisons. Note that all the aforementioned enhancements are not part of the original designs of the baseline methods. However, we empirically find that they significantly improve baseline performance.

MMLUPro [46]、GPQA [47]、FinQA [48]和 MedQA [49])，以及情感类 (EmoryNLP [50]和 MELD [51])。此外，为了评估 Avengers 在分布外 (OOD) 任务上的泛化能力，我们进一步为每个类别额外包含一个数据集：MathBench [52]、StudentEval [53]、Winogrande [54]、BRAINTEASER [55]和 DailyDialog [56]。请注意，这额外的五个数据集仅用于评估 OOD 泛化性能。关于这些数据集的详细信息见附录 A.3 和表 9。

复仇者联盟的实现 我们统一采用 gte-qwen2-7B-instruct 作为嵌入模型，并使用经典的 K-Means 算法，将聚类数量设置为 $K = 64$ 。每个数据集被随机分为 70%用于拟合聚类中心，其余 30%用于性能评估。最初，我们考虑一个由 22 个开源候选 LLM (约 7B) 组成的池，从中自动选择 $m = 10$ 个模型。在推理过程中，我们默认采用自一致性 (SC) 策略，将采样轮数设置为 10。所有实验均使用五个随机种子 (42、999、2024、2025 和 3407) 重复五次，并报告平均结果。

详细的模型信息和部署细节可在附录 A.2 中找到。

基线方法 我们从两个不同的技术方向 (基于混合和基于路由) 选取基线方法，以全面评估 Avengers 的有效性。此外，为了提供直观的性能基准，我们纳入了三个专有模型：gpt-4o-2024-08-06 (GPT-4o)、gpt-4.5-preview (GPT-4.5-preview) 和 gpt-4.1-2025-04-14 (GPT-4.1)。具体而言，我们考虑以下基线方法：

- 随机路由器：对于每个查询，从候选集 S 中随机选择一个模型。
- LLM 路由器：一种基于提示的路由器，利用 LLM (Qwen2.5-7B-Instruct) 根据总结候选模型性能的自然语言配置文件来选择模型。
- RouterDC [8]：利用双重对比学习来训练查询和模型的嵌入。对比目标确保查询嵌入接近合适的模型，并且语义相似的查询能够聚集在一起。
- EmbedLLM [9]：一种基于嵌入的路由器，通过二元交叉熵损失训练，用于预测模型与查询的兼容性。
- MODEL-SAT [10]：采用能力指令微调，将候选模型的性能转化为文本能力描述，这些描述被嵌入并传递给可训练的 LLM，以预测每个查询的模型适用性。
- 混合智能体 (MoA) [11]：一种分层混合方法，其中多层 LLM 迭代地细化或聚合候选答案，最终通过一个聚合步骤完成。
- 符号化混合专家 (SMoE) [14]：一种使用技能关键词匹配的混合方法。根据查询派生的技能关键词选择候选模型，并通过单层 MoA 聚合其输出。

这些基线方法与 Avengers 之间的方法论差异总结在表 1 中。更详细的实现细节可在附录 A.4 中找到。

对基线的增强值得注意的是，为了更好地使基线方法与我们的实验场景对齐并促进全面比较，我们对某些基线引入了有针对性的增强。所有基线方法均使用由 Avengers 框架自动选择的同一组 10 个模型。具体而言，对于基于混合的方法 MoA，我们额外评估了一个 oracle 变体——MoA (Oracle)，在该变体中，我们手动为每个任务选择表现最好的前 3 个模型，而非使用全部 10 个模型。这种方法解决了原始 MoA 设置中因上下文窗口过长而导致的性能下降问题。对于基于可训练路由器的方法 (RouterDC、EmbedLLM 和 MODEL-SAT)，我们不采用固定的训练步数作为停止标准，而是报告每种方法在测试集上达到峰值时的性能。此外，表 2 和表 3 中报告的所有 SLM 和基于路由器的结果均采用了与 Avengers 相同的自一致性 (SC) 策略，即每个查询采样 10 次，以确保比较的公平性和稳健性。请注意，上述所有增强均不属于基线方法的原始设计。然而，我们通过实验发现，它们显著提升了基线性能。

Table 1: Methodological comparison between baseline methods and the *Avengers*. **Trainable NN**: additional neural components requiring training. **Arch.**: effort in architecture design. **Prompt Eng.**: effort in prompt design. **New Task**: how to adapt to new tasks. **New Model**: how to adapt to new models. **Model Scale**: can scale with the number of SLMs. **Auto Init. Sel.**: can automatically select the initial SLM subset.

Method	Trainable NN [†]	Arch.	Prompt	New Task [‡]	New Model [‡]	Model Scale [§]	Auto Init. Sel.
RouterDC [8]	Emb+MLP	Free	Free	LS+Retrain NN		✓	✗
EmbedLLM [9]	MLP	Free	Free	LS+Retrain NN		✓	✗
MODEL-SAT [10]	Emb+MLP+SLM	Free	Hand-Craft	LS+Retrain NN		✓	✗
MoA [11]	Free	Hand-Craft	Hand-Craft	Free		✗	✗
Symb.-MoE [14]	Free	Free	Hand-Craft	LS+Regen Profile		✓	✗
Avenger (ours)	Free	Free	Free	LS+Recluster	LS	✓	✓

[†] Emb = trainable embedding model; MLP = small feed-forward head. [‡] LS = labeled samples.

[§] Mixture of Agents exceeds the maximum context window for SLMs (e.g., 8192 tokens for Gemma-2-9B) as the number of SLMs increases.

4.2 The Avengers Achieves State-Of-The-Art Performance

Table 2 presents a detailed performance comparison between the *Avengers* and three groups of baselines across 15 datasets in five task categories. First, the *Avengers* **achieves an average score of 70.54, surpassing GPT-4.1’s 69.20**, demonstrating that dynamically selecting among smaller models yields performance comparable to large-scale single models. Specifically, the *Avengers* outperforms GPT-4.1 on 10 of the 15 datasets, achieving an average improvement of 18.21% on mathematical tasks, with notable improvements of 31.86% on MBPP and 15.41% on KORBench. However, GPT-4.1 retains its advantage on specialized knowledge-intensive tasks (e.g., GPQA, MedQA) due to its larger parameter size and broader knowledge base. Nevertheless, the *Avengers* exceeds the max expert by 6.89% on MMLUPro (61.94 to 66.21), nearly matching GPT-4.1, underscoring the complementary potential of smaller models.

Second, mixture-based methods (MoA, SMoE, and MoA (Oracle)) struggle under our setting. Under identical candidate model pools, router-based methods generally outperform mixture-based methods. This is primarily due to limited model context windows and weaker instruction-following abilities of smaller models, resulting in lower performance (< 61) compared to the *Avengers*.

Third, even when we enhance router-based methods by selecting their peak on test sets, the *Avengers* still **achieves the best average performance** without involving additional training parameters. Moreover, traditional router-based methods often struggle in Out-of-Distribution (OOD) settings. To compare their generalization ability, we introduce a new dataset for each of the five task categories to evaluate performance. As shown in Table 3, the *Avengers* **exhibits the most robust generalization** among all compared methods, achieving an average score of 74.42. This performance surpasses the best router-based baseline by EmbedLLM (+8.14%).

It should be noted that the above baselines results have been enhanced with techniques mentioned in the last section. We choose to report results with these enhancements, rather than strictly adhering to the original implementations, for two reasons: (i) to isolate the impact of the routing mechanism itself, enabling a fairer comparison between our clustering-based approach and other routing methods; and (ii) to contribute constructively to the field by demonstrating that some of the performance gains achieved by the *Avengers* can, in principle, be reproduced by other routing methods under similar improvements, but with additional, lightweight operations introduced by the *Avengers*.

4.3 The Key Elements that Make the Avengers Work

To understand why the *Avengers* is both simple and effective, we conduct targeted ablations on four design choices—**embedding model**, **clustering method**, **model selection**, and **ensemble strategy**—while keeping all other components fixed.

Embedding Model Since clustering quality directly depends on the embedding model’s semantic representation, we investigate how the choice of embedding model affects the performance of the

表 1: 基线方法与 Avengers 之间的方法论对比。可训练神经网络: 需要训练的额外神经组件。架构设计: 架构设计中的工作量。提示工程: 提示设计中的工作量。新任务: 如何适应新任务。新模型: 如何适应新模型。模型规模: 能否随 SLM 数量扩展。自动初始选择: 能否自动选择初始 SLM 子集。

方法	可训练	神经网络架构	提示	新任务	新模型	模型规模	自动初始化	选择							
RouterDC [8]	嵌入+MLP	自由	自由	标签平滑+重训练	神经网络	✓	X	EmbedLLM [9]	MLP	自由	自由	标签平滑+重训练	神经网络	✓	X
MODEL-SAT [10]	嵌入+MLP+SLM	自由	手工设计	标签平滑+重训练	神经网络	✓	X								
MoA [11]	免费手工	手工	免费	X	X	Symb.-MoE [14]	免费	免费手工	LS+Regen	配置文件	✓	X			
Avenger（我们的）	免费	免费	免费	LS+重新聚类	LS	✓	✓								

[†] Emb = 可训练嵌入模型; MLP = 小型前馈头。LS = 带标签样本。
[§] 混合智能体 (Mixture of Agents) 随着 SLM 数量的增加, 会超出 SLM 的最大上下文窗口 (例如, Gemma-2-9B 的 8192 个 token)。

4.2 复仇者联盟实现最先进性能

表 2 展示了复仇者联盟与三组基线模型在五个任务类别、15 个数据集上的详细性能对比。首先, 复仇者联盟取得了 70.54 的平均得分, 超越了 GPT-4.1 的 69.20, 这表明在较小模型中动态选择能够产生与大规模单模型相当的性能。具体而言, 复仇者联盟在 15 个数据集中的 10 个上优于 GPT-4.1, 在数学任务上实现了 18.21% 的平均提升, 其中在 MBPP 上提升了 31.86%, 在 KORBench 上提升了 15.41%。然而, GPT-4.1 凭借其更大的参数量和更广泛的知识库, 在专业知识密集型任务 (如 GPQA、MedQA) 上仍保持优势。尽管如此, 复仇者联盟在 MMLUPro 上超出最佳专家模型 6.89% (从 61.94 提升至 66.21), 几乎与 GPT-4.1 持平, 这凸显了较小模型之间的互补潜力。

其次, 基于混合的方法 (MoA、SMoE 和 MoA (Oracle)) 在我们的设置下表现不佳。在相同的候选模型池中, 基于路由的方法通常优于基于混合的方法。这主要是由于较小的模型上下文窗口有限且指令遵循能力较弱, 导致其性能 (<61) 低于 Avengers。

第三, 即使我们通过选择测试集上的峰值来增强基于路由的方法, Avengers 在不引入额外训练参数的情况下仍能实现最佳平均性能。此外, 传统的基于路由的方法在分布外 (OOD) 设置中往往表现不佳。为了比较它们的泛化能力, 我们为五个任务类别各引入了一个新数据集来评估性能。如表 3 所示, Avengers 在所有比较方法中展现出最强的泛化能力, 平均得分为 74.42。这一性能超越了最佳基于路由的基线方法 EmbedLLM (+8.14%)。

需要注意的是, 上述基线结果已通过上一节提到的技术进行了增强。我们选择报告这些增强后的结果, 而非严格遵循原始实现, 原因有二: i) 为了隔离路由机制本身的影响, 从而在我们基于聚类的方法与其他路由方法之间实现更公平的比较; ii) 为了对该领域做出建设性贡献, 证明复仇者联盟所取得的部分性能提升, 原则上可以通过其他路由方法在类似改进下复现, 但复仇者联盟引入了额外的轻量级操作。

4.3 使复仇者联盟奏效的关键要素

为了理解为什么“复仇者联盟”既简单又有效, 我们在保持其他组件不变的情况下, 对四个设计选择进行了有针对性的消融实验——嵌入模型、聚类方法、模型选择和集成策略。

嵌入模型 由于聚类质量直接依赖于嵌入模型的语义表示, 我们研究了嵌入模型的选择如何影响“复仇者联盟”的性能。

Table 2: Comparison of the *Avengers* with baselines. *Oracle** represents the best achievable score by selecting the optimal model per query. [†]Peak score on the test set. Max Expert represents the best performance of the ten models on the current dataset.

Setting	Mathematics			Code		Logical			Knowledge				Affective		Avg.	
	AIME	M500	LMB	MBPP	HE	KOR	K&K	BBH	ARCC	MP	GPQA	FinQA	MedQA	Emory		MELD
Small Language Model (Enhanced)																
DS-Qwen	61.67	93.00	65.71	56.67	43.90	52.00	76.10	80.00	88.82	59.14	59.15	68.53	40.85	29.84	38.64	60.93
Fin-R1	13.33	80.20	35.00	69.30	77.44	37.52	18.14	68.24	91.89	51.55	20.76	70.71	67.09	41.18	57.55	53.33
Qwen-it	15.00	78.60	37.14	70.73	81.10	49.60	35.92	63.24	88.91	58.34	30.13	68.00	65.28	39.89	57.87	55.98
Qwen-Coder	11.67	75.60	35.00	76.39	78.66	34.64	27.57	61.48	86.52	49.25	32.14	65.13	55.38	40.46	59.58	52.63
gemma-2-it	1.67	54.00	22.14	62.32	64.02	34.32	15.71	63.52	89.42	53.55	34.38	66.26	66.46	39.60	54.38	48.12
glm-4-chat	5.00	58.40	22.14	62.01	65.85	37.60	21.57	47.59	92.15	51.75	31.25	58.59	64.96	41.32	57.46	47.84
Llama-3.1-it	1.67	49.80	20.71	61.81	71.95	27.60	11.71	65.74	88.48	47.35	25.67	53.97	69.76	35.15	51.62	45.53
Granite-3.1-it	1.67	61.40	20.00	37.17	39.63	31.44	19.29	36.39	85.24	44.06	34.82	65.74	59.70	39.60	48.30	41.63
UltraMedical	1.67	76.20	15.71	52.77	58.54	18.72	20.14	40.46	85.67	43.76	19.87	60.68	72.90	31.28	43.34	42.78
cogito-v1	3.33	59.60	22.86	51.33	70.12	44.48	28.00	75.83	90.01	61.94	31.92	64.69	69.36	39.02	55.68	51.21
Average	11.67	68.68	29.64	60.05	65.12	36.79	27.42	60.25	88.71	52.07	32.01	64.23	63.17	37.73	52.44	50.00
Max Expert	61.67	93.00	65.71	76.39	81.10	52.00	76.10	80.00	92.15	61.94	59.15	70.88	72.90	41.32	59.58	69.59
Oracle*	63.33	96.60	69.29	89.43	95.73	65.76	82.43	96.85	96.83	86.41	85.49	85.44	91.99	66.86	82.71	83.68
Proprietary Model Baseline																
GPT-4o	10.00	76.00	35.71	82.64	85.36	57.68	32.57	79.53	93.86	59.84	44.42	72.28	82.17	38.31	52.92	60.22
GPT-4.5-Preview	31.67	88.60	50.00	86.69	69.51	57.44	89.29	91.36	94.60	59.94	51.78	71.75	92.85	39.45	47.84	68.18
GPT-4.1	45.00	89.20	52.14	57.70	92.07	48.48	84.43	90.74	95.39	65.13	62.05	71.05	88.77	39.45	56.33	69.20
Mixture-based Baseline																
MoA	23.33	82.67	33.33	54.61	84.51	49.87	34.28	47.53	91.81	64.54	50.37	71.04	71.20	39.17	52.76	56.70
MoA (Oracle)	27.77	84.67	54.76	69.40	83.54	51.28	31.14	57.50	92.75	61.74	50.44	70.71	76.05	39.45	51.95	60.21
SMoE	43.00	88.20	42.86	65.81	70.12	48.88	43.71	67.69	87.63	58.94	43.30	65.21	64.96	37.30	51.38	58.60
Router-based Baseline (Enhanced)																
Random Router	10.00	66.20	32.86	61.29	62.80	35.60	27.57	61.20	88.65	51.25	33.04	64.60	62.61	38.16	50.89	49.78
LLM Router	61.67	91.80	64.29	69.92	78.05	37.12	74.71	80.28	88.99	60.14	45.31	66.78	73.84	32.71	45.78	64.76
RouterDC [†]	58.89	89.87	64.76	68.53	73.20	44.64	73.33	73.33	90.00	54.68	43.41	68.12	66.81	42.00	58.87	64.70
MODEL-SAT [†]	61.11	91.73	65.24	69.42	80.00	54.77	72.76	82.96	90.34	65.12	58.67	68.70	66.70	40.19	58.00	68.38
EmbedLLM [†]	61.11	90.80	64.76	74.27	78.80	55.73	73.81	85.06	91.14	65.32	57.33	67.12	71.83	41.33	60.22	69.24
The Avengers (ours)	61.67	92.89	65.71	76.08	84.86	55.95	76.14	84.07	92.39	66.21	55.76	71.53	73.89	41.38	59.61	70.54
- vs GPT-4.1 (%)	↑37.04	↑14.14	↑26.03	↑31.86	↑7.83	↑15.41	↑9.82	↑7.35	↑3.15	↑1.66	↑10.14	↑0.67	↑16.76	↑4.90	↑5.83	↑1.95
- vs MoA (Oracle) (%)	↑122.06	↑9.71	↑20.00	↑9.63	↑1.58	↑9.11	↑144.52	↑46.21	↓0.39	↑7.24	↑10.54	↑1.16	↑2.84	↑4.90	↑14.75	↑17.16
- vs EmbedLLM (%)	↑0.91	↑2.31	↑1.47	↑2.44	↑7.69	↑0.39	↑3.16	↑1.16	↑1.37	↑1.37	↑2.75	↑6.57	↑2.87	↑0.12	↑1.00	↑1.88

Table 3: Out-of-Distribution performance. [†]Peak score on the test set.

Setting	MathBench	StudentEval	Winogrande	BrainTeaser	DailyDialog	Avg.
<i>Small Language Model (Enhanced)</i>						
DS-Qwen	97.33	53.83	65.98	68.46	38.30	64.78
Fin-R1	65.33	65.69	77.82	70.13	43.30	64.45
Qwen-it	83.33	65.48	69.85	67.78	53.30	67.95
Qwen-Coder	75.33	63.80	67.56	59.84	57.70	64.85
gemma-2-it	53.30	64.22	73.16	76.00	50.90	63.52
glm-4-chat	56.67	60.65	77.74	62.68	59.10	63.37
Llama-3.1-it	49.33	68.84	64.09	70.03	37.70	58.00
Granite-3.1-it	57.33	40.50	72.77	61.80	35.20	53.52
UltraMedical	38.67	57.61	62.43	62.88	32.60	50.84
cogito-v1	66.00	56.45	64.09	73.16	37.70	59.48
Average	64.26	59.71	69.55	67.28	44.58	61.07
Max Expert	97.33	68.84	77.82	76.00	59.10	67.95
Oracle	100.00	84.89	95.26	94.32	75.30	89.95
<i>Router Baseline (Enhanced)</i>						
Random Router	64.67	55.82	69.53	66.70	43.60	60.06
LLM Router	94.67	67.79	65.27	67.48	37.10	66.46
RouterDC [†]	77.27	62.50	69.77	70.91	59.20	66.89
MODEL-SAT [†]	97.20	58.22	68.37	70.58	48.40	68.55
EmbedLLM [†]	97.20	63.06	71.11	69.34	47.98	69.77
The Avengers (ours)	96.67	68.52	77.43	71.79	57.70	74.42
- vs EmbedLLM (%)	↓0.55	↑4.98	↑19.17	↑2.20	↑24.30	↑8.14

Avengers. Specifically, we compare five embedding models spanning different parameter sizes and embedding dimensions (Table 4), including **bge-m3** [57], **text-embedding-3-small** (text-3-small) [58], **gte-qwen2-1.5B-instruct** (gte-q2-1.5B) [59], **text-embedding-3-large** (text-3-large) [58]

我们研究了嵌入模型的选择如何影响表 2 中 Avengers 与基线模型对比的性能。Oracle 表示通过为每个查询选择最优模型所能达到的最佳得分。测试集上的峰值得分。Max Expert 表示当前数据集上十个模型中的最佳性能

设置	数学 代码 逻辑 知识 情感	平均
	AIME M500。LMB。MBPP HE。KOR。K&K。BBH ARCC MP。GPQA FinQA MedQA Emory。MELD	
小型语言模型（增强版）		
DS-Qwen 61.67 93.00 65.71 56.67 43.90 52.00 76.10 80.00 88.82 59.14 59.15 68.53 40.85 29.84 38.64 60.93 Fin-R1 13.33 80.20 35.00 69.30 77.44 37.52 18.14 68.24 91.89 51.55 20.76 70.71 67.09 41.18 57.55 53.33 Qwen-it 15.00 78.60 37.14 70.73 81.10 49.60 35.92 63.24 88.91 58.34 30.13 68.00 65.28 39.89 57.87 55.98 Qwen-Coder 11.67 75.60 35.00 76.39 78.66 34.64 27.57 61.48 86.52 49.25 32.14 65.13 55.38 40.46 59.58 52.63 gemma-2-it 1.67 54.00 22.14 62.32 64.02 34.32 15.71 63.52 89.42 53.55 34.38 66.26 66.46 39.60 54.38 48.12 glm-4-chat 5.00 58.40 22.14 62.01 65.85 37.60 21.57 47.59 92.15 51.75 31.25 58.59 64.96 41.32 57.46 47.84 Llama-3.1-it 1.67 49.80 20.71 61.81 71.95 27.60 11.71 65.74 88.48 47.35 25.67 53.97 69.76 35.15 51.62 45.53 Granite-3.1-it 1.67 61.40 20.00 37.17 39.63 31.44 19.29 36.39 85.24 44.06 34.82 65.74 59.70 39.60 48.30 41.63 UltraMedical 1.67 76.20 15.71 52.77 58.54 18.72 20.14 40.46 85.67 43.76 19.87 60.68 72.90 31.28 43.34 42.78 cogito-v1 3.33 59.60 22.86 51.33 70.12 44.48 28.00 75.83 90.01 61.94 31.92 64.69 69.36 39.02 55.68 51.21		
平均 11.67 68.68 29.64 60.05 65.12 36.79 27.42 60.25 88.71 52.07 32.01 64.23 63.17 37.73 52.44 50.00 最大专家 61.67 93.00 65.71 76.39 81.10 52.00 76.10 80.00 92.15 61.94 59.15 70.88 72.90 41.32 59.58 69.59 神谕 63.33 96.60 69.29 89.43 95.73 65.76 82.43 96.85 96.83 86.41 85.49 85.44 91.99 66.86 82.71 83.68		
专有模型基线		
GPT-4o 10.00 76.00 35.71 82.64 85.36 57.68 32.57 79.53 93.86 59.84 44.42 72.28 82.17 38.31 52.92 60.22 GPT-4.5-Preview 31.67 88.60 50.00 86.69 69.51 57.44 89.29 91.36 94.60 59.94 51.78 71.75 92.85 39.45 47.84 68.18 GPT-4.1 45.00 89.20 52.14 57.70 92.07 48.48 84.43 90.74 95.39 65.13 62.05 71.05 88.77 39.45 56.33 69.20		
基于混合的基线		
MoA 23.33 82.67 33.33 54.61 84.51 49.87 34.28 47.53 91.81 64.54 50.37 71.04 71.20 39.17 52.76 56.70 MoA (Oracle) 27.77 84.67 54.76 69.40 83.54 51.28 31.14 57.50 92.75 61.74 50.44 70.71 76.05 39.45 51.95 60.21 SMoE 43.00 88.20 42.86 65.81 70.12 48.88 43.71 67.69 87.63 58.94 43.30 65.21 64.96 37.30 51.38 58.60		
基于路由器的基线（增强版）		
随机路由器 10.00 66.20 32.86 61.29 62.80 35.60 27.57 61.20 88.65 51.25 33.04 64.60 62.61 38.16 50.89 49.78 LLM 路由器 61.67 91.80 64.29 69.92 78.05 37.12 74.71 80.28 88.99 60.14 45.31 66.78 73.84 32.71 45.78 64.76 RouterDC 58.89 89.87 64.76 68.53 73.20 44.64 73.33 73.33 90.00 54.68 43.41 68.12 66.81 42.00 58.87 64.70 MODEL-SAT 61.11 91.73 65.24 69.42 80.00 54.77 72.76 82.96 90.34 65.12 58.67 68.70 66.70 40.19 58.00 68.38 EmbedLLM 61.11 90.80 64.76 74.27 78.80 55.73 73.81 85.06 91.14 65.32 57.33 67.12 71.83 41.33 60.22 69.24		
复仇者联盟（我们的方法） 61.67 92.89 65.71 76.08 84.86 55.95 76.14 84.07 92.39 66.21 55.76 71.53 73.89 41.38 59.61 70.54		
- 对比 GPT-4.1 (%) ↑ 37.04 ↑ 4.14 ↑ 26.03 ↑ 31.86 ↓ 7.83 ↑ 15.41 ↓ 9.82 ↓ 7.35 ↓ 3.15 ↑ 1.66 ↓ 10.14 ↑ 0.67 ↓ 16.76 ↑ 4.90 ↑ 5.83 ↑ 1.95		
- 对比 MoA (Oracle) (%) ↑ 122.06 ↑ 9.71 ↑ 20.00 ↑ 9.63 ↑ 1.58 ↑ 9.11 ↑ 144.52 ↑ 46.21 ↓ 0.39 ↑ 7.24 ↑ 10.54 ↑ 1.16 ↓ 2.84 ↑ 4.90 ↑ 14.75 ↑ 17.16		
- 对比 EmbedLLM (%) ↑ 0.91 ↑ 2.31 ↑ 1.47 ↑ 2.44 ↑ 7.69 ↑ 0.39 ↑ 3.16 ↓ 1.16 ↑ 1.37 ↑ 1.37 ↓ 2.75 ↑ 6.57 ↑ 2.87 ↑ 0.12 ↓ 1.00 ↑ 1.88 表 3: 分布外性能。测试集上的峰值分数。		

设置	MathBench StudentEval Winogrande BrainTeaser DailyDialog 平均
小型语言模型（增强版）	
DS-Qwen 97.33 53.83 65.98 68.46 38.30 64.78	
Fin-R1 65.33 65.69 77.82 70.13 43.30 64.45	
Qwen-it 83.33 65.48 69.85 67.78 53.30 67.95 Qwen-Coder 75.33 63.80 67.56 59.84 57.70 64.85 gemma-2-it 53.30 64.22 73.16 76.00 50.90 63.52 glm-4-chat 56.67 60.65 77.74 62.68 59.10 63.37 Llama-3.1-it 49.33 68.84 64.09 70.03 37.70 58.00 Granite-3.1-it 57.33 40.50 72.77 61.80 35.20 53.52 UltraMedical 38.67 57.61 62.43 62.88 32.60 50.84 cogito-v1 66.00 56.45 64.09 73.16 37.70 59.48	
平均值 64.26 59.71 69.55 67.28 44.58 61.07	
Max Expert 97.33 68.84 77.82 76.00 59.10 67.95	
Oracle 100.00 84.89 95.26 94.32 75.30 89.95 路由器基线（增强）	
随机路由器 64.67 55.82 69.53 66.70 43.60 60.06 LLM 路由器 94.67 67.79 65.27 67.48 37.10 66.46 RouterDC 77.27 62.50 69.77 70.91 59.20 66.89 MODEL-SAT 97.20 58.22 68.37 70.58 48.40 68.55 EmbedLLM 97.20 63.06 71.11 69.34 47.98 69.77	
复仇者联盟（我们的） 96.67 68.52 77.43 71.79 57.70 74.42	
- 对比 EmbedLLM (%) ↓ 0.55 ↑ 4.98 ↑ 19.17 ↑ 2.20 ↑ 24.30 ↑ 8.14	

复仇者联盟。具体来说，我们比较了五个覆盖不同参数规模和嵌入维度的嵌入模型（表 4），包括 bge-m3 [57]、text-embedding-3-small (text-3small) [58]、gte-qwen2-1.5B-instruct (gte-q2-1.5B) [59]、text-embedding-3-large (text-3-large) [59]、

Table 4: Average score of the *Avengers* when equipped with different embedding models.

Embed.	Para./Dim.	Avg.
bge-m3	0.56/1 024	70.28
text-3-small [†]	-/1 536	70.42
gte-q2-1.5B	1.5/1 536	70.48
text-3-large [†]	-/3 072	70.70
gte-q2-7B	7/3 584	<u>70.54</u>

Table 5: Comparison of average scores for different clustering methods.

Cluster Method	Avg.
K-Means	70.54
Hierarchical Clustering	70.71
GMM	70.58
Spectral Clustering	<u>70.61</u>
BIRCH	<u>70.04</u>

Table 6: Ablation on the *Avengers* ensemble strategies (excluding code tasks).

Ensemble Strategy	Avg.
Direct (CoT)	59.19
Aggregation	56.73
Model-Switch	<u>66.42</u>
Self-Consistency	66.98

and **gte-qwen2-7B-instruct** (gte-q2-7B) [59]. Results show minimal performance variation (70.28 $\rightarrow$ 70.70), demonstrating that the proposed method is robust to changes in embedding model selection.

Clustering Method Since the *Avengers* relies on query clustering, the choice of clustering method could potentially impact overall performance. To assess this influence, we conduct experiments using five classical clustering algorithms: **K-Means** [15, 16], **Hierarchical Clustering** [17], **Gaussian Mixture Models (GMM)** [18], **Spectral Clustering** [19] and **BIRCH** [20]. For all experiments, we fix the number of clusters at $K = 64$ (details in Appendix B.2). As shown in Table 5, performance varies minimally (70.04–70.71), indicating that the *Avengers* framework is robust and stable across different clustering algorithms.

Model Selection Figure 3 shows the performance as the number of selected models in the *Avengers* increases from 1 to 22. The overall performance rises sharply, matching GPT-4.1 with only 3 models, and plateaus at approximately 10 models. Knowledge, code, and affective tasks exhibit similar trends, while mathematics and logic tasks gain little due to a single model’s dominance. These results confirm that the *Avengers* effectively leverages model complementarity, achieving strong general-purpose performance without needing to deploy all available models.

Ensemble Strategy To analyze ensemble strategies within the *Avengers*, we compare Direct (CoT), Aggregation, Model-Switch, and Self-Consistency. Among these, Self-Consistency achieves the best overall performance due to its simplicity, effectiveness, and robustness across tasks.

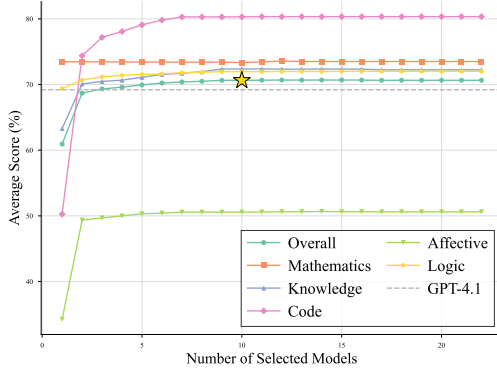


Figure 3: Impact of selected models’ number.

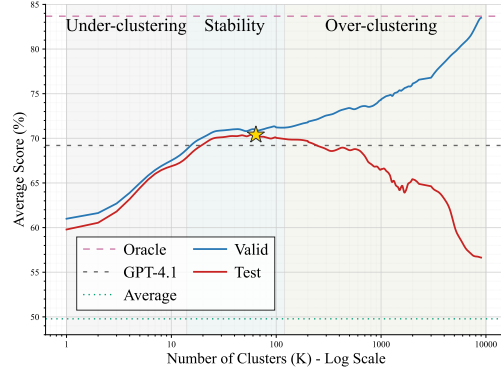


Figure 4: Impact of cluster count K .

The Impact of Cluster Count K Figure 4 demonstrates the performance dynamics of the *Avengers* on both validation and test sets as the cluster count K varies from 1 to 10,000 (logarithmic scale on the horizontal axis). Notably, our method surpasses GPT-4.1 across a wide range of K values (approximately 14 to 140), indicating robustness and minimal sensitivity to the choice of K .

5 Conclusion

In this paper, we introduce the *Avengers*, a simple yet effective framework to unite multiple smaller language models (SLMs) and challenge the dominance of proprietary large models. The core of

text-embedding-3-large
(text-3-large) [58] 表 4: 配

嵌入 参数/维度 平均
bge-m3 0.56/1 024 70.28 text-3- small-/1 536 70.42 gte-q2-1.5B 1.5/1 536 70.48 text-3-large-/3 072 70.70 gte- q2-7B 7/3 584 70.54

表 5: 不同聚类方法的平均
得分比较。

聚类方法 平均分
K-Means 70.54 层次聚类 70.71 GMM 70.58 谱聚类 70.61 BIRCH 70.04

表 6: 复仇者联盟集成策略的
消融实验 (不包括代码任务)

集成策略平均值
直接 (思维链) 59.19 聚合 56.73 模型切换 66.42 自一致性 66.98

以及 gte-qwen2-7B-instruct (gte-q2-7B) [59]。结果显示性能变化极小 (70.28 $\rightarrow$ 70.70)，表明所提出的方法对嵌入模型选择的变化具有鲁棒性。

聚类方法 由于 Avengers 依赖于查询聚类，聚类方法的选择可能对整体性能产生影响。为评估这一影响，我们使用五种经典聚类算法进行实验：K-Means [15, 16]、层次聚类 [17]、高斯混合模型 (GMM) [18]、谱聚类 [19] 和 BIRCH [20]。在所有实验中，我们将聚类数量固定为 $K = 64$ (详见附录 B.2)。如表 5 所示，性能变化极小 (70.04–70.71)，表明 Avengers 框架在不同聚类算法下具有鲁棒性和稳定性。

模型选择 图 3 展示了随着复仇者联盟中选定模型数量从 1 增加到 22 时的性能变化。整体性能急剧上升，仅用 3 个模型即可媲美 GPT-4.1，并在约 10 个模型时趋于平稳。知识、代码和情感类任务表现出相似趋势，而数学和逻辑类任务因单一模型的优势而提升有限。这些结果证实，复仇者联盟有效利用了模型的互补性，无需部署所有可用模型即可实现强大的通用性能。

集成策略 为了分析复仇者联盟中的集成策略，我们比较了直接 (思维链)、聚合、模型切换和自一致性。其中，自一致性凭借其简单性、有效性和跨任务的稳健性，取得了最佳的整体性能。

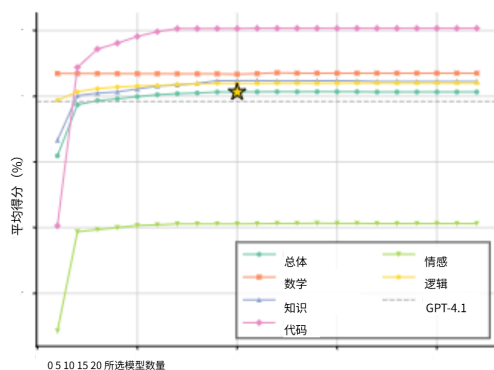


图 3: 所选模型数量的影响。

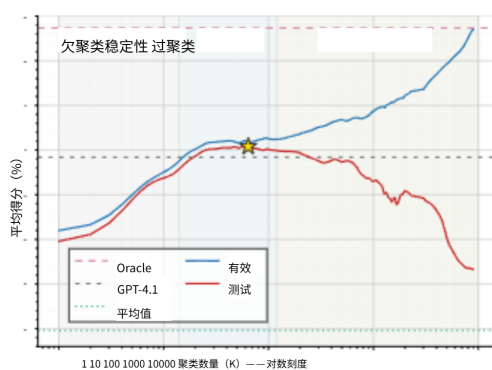


图 4: 聚类数量 K 的影响。

聚类数量 K 的影响 图 4 展示了当聚类数量 K 从 1 变化到 10,000 时 (横轴为对数刻度)，Avengers 在验证集和测试集上的性能动态。值得注意的是，我们的方法在广泛的 K 值范围内 (约 14 至 140) 超越了 GPT-4.1，这表明该方法具有鲁棒性，且对 K 的选择敏感性极低。

5 结论

在本文中，我们介绍了 Avengers，一个简单而有效的框架，旨在联合多个小型语言模型 (SLMs)，挑战专有大型模型的主导地位。其核心在于

the *Avengers* involves straightforward embedding, clustering, scoring, and voting, without requiring neural network training, prompt engineering, or careful architecture-specific model choices. Moreover, it is highly adaptable to new domains and new models, and supports automatic model set construction.

Experimental results show that by combining 10 SLMs (each with approximately 7B parameters), the *Avengers* achieves highly competitive performance across diverse tasks. Specifically, it surpasses OpenAI’s flagship model, GPT-4.1, on nine out of 15 diverse datasets spanning mathematics, coding, logical reasoning, knowledge, and affective recognition tasks. Notably, it achieves an 18.21% improvement over GPT-4.1 on mathematics tasks and a 7.46% improvement on coding tasks. Additionally, the *Avengers* demonstrates superior out-of-distribution (OOD) generalization, outperforming state-of-the-art router-based methods by at least 8%. Further analyses reveal the robustness of the *Avengers* across different clustering methods, embedding model choices, and key hyperparameters such as the number of clusters (k). With an increasing number of integrated models, the *Avengers* exhibits complementary advantages, especially in knowledge, code and affective tasks.

The superior performance of the *Avengers* highlights the untapped potential of community-driven, open-source models and provides a promising pathway toward democratizing the advancements of LM research.

References

- [1] Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, et al. Fin-r1: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv:2503.16252*, 2025.
- [2] Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.
- [3] Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Binqing Qi, Xuekai Zhu, et al. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081, 2024.
- [4] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.
- [5] Michelle Flores Vryn and Meena Das. Building community-centered ai collaborations. *Stanford Social Innovation Review*, 2025. doi: 10.48558/KF5R-C507. URL <https://ssir.org/articles/entry/community-centered-ai-collaborations>.
- [6] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.
- [7] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, 2023.
- [8] Shuhao Chen, Weisen Jiang, Baijiong Lin, James Kwok, and Yu Zhang. Routerdc: Query-based router by dual contrastive learning for assembling large language models. *Advances in Neural Information Processing Systems*, 37:66305–66328, 2024.
- [9] Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, and Kannan Ramchandran. Embedllm: Learning compact representations of large language models. *arXiv preprint arXiv:2410.02223*, 2024.
- [10] Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Capability instruction tuning: A new paradigm for dynamic llm routing. *arXiv preprint arXiv:2502.17282*, 2025.
- [11] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

复仇者联盟的核心涉及简单的嵌入、聚类、评分和投票，无需神经网络训练、提示工程或针对特定架构的精细模型选择。此外，它对新领域和新模型具有高度适应性，并支持自动模型集构建。

实验结果表明，通过组合 10 个小型语言模型（每个约 70 亿参数），Avengers 在多种任务中展现出极具竞争力的性能。具体而言，在涵盖数学、编程、逻辑推理、知识及情感识别任务的 15 个多样化数据集中，它在其中 9 个上超越了 OpenAI 的旗舰模型 GPT-4.1。值得注意的是，在数学任务上，它相较于 GPT-4.1 实现了 18.21% 的提升，在编程任务上则提升了 7.46%。此外，Avengers 展现出优越的分布外（OOD）泛化能力，比最先进的基于路由的方法至少高出 8%。进一步分析揭示了 Avengers 在不同聚类方法、嵌入模型选择以及关键超参数（如聚类数 k ）下的稳健性。随着集成模型数量的增加，Avengers 展现出互补优势，尤其在知识、代码和情感任务中表现突出。

Avengers 的卓越性能凸显了社区驱动、开源模型尚未被充分挖掘的潜力，并为推动语言模型研究进展的民主化提供了一条有前景的路径。

参考文献

- [1] 刘兆伟, 郭鑫, 楼方琦, 曾凌峰, 牛金一, 王子轩, 徐佳杰, 蔡伟格, 杨子薇, 赵雪倩, 等. Fin-r1: 一种通过强化学习进行金融推理的大型语言模型. arXiv 预印本 arXiv:2503.16252, 2025 年。
- [2] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, 等. Qwen2.5-coder 技术报告. arXiv 预印本 arXiv:2409.12186, 2024.
- [3] Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, 等. Ultramedical: 构建生物医学领域的专业通才. 神经信息处理系统进展, 37: 26045–26081, 2024.
- [4] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, 等. 通过隐式奖励进行过程强化. arXiv 预印本 arXiv:2502.01456, 2025.
- [5] Michelle Flores Vryn 和 Meena Das. 构建以社区为中心的人工智能合作. 斯坦福社会创新评论, 2025. doi: 10.48558/KF5R-C507. URL <https://ssir.org/articles/entry/community-centered-ai-collaborations>.
- [6] Roy Schwartz, Jesse Dodge, Noah A Smith 和 Oren Etzioni. 绿色人工智能. 《美国计算机协会通讯》, 63(12): 54–63, 2020 年。
- [7] Dongfu Jiang, Xiang Ren 和 Bill Yuchen Lin. LLM-Blender: 通过成对排序和生成式融合集成大型语言模型. 收录于第 61 届年会论文集
计算语言学协会会议论文集（第 1 卷：长文），第 14165–14178 页，2023 年。
- [8] Shuhao Chen, Weisen Jiang, Baijiong Lin, James Kwok, 和 Yu Zhang. Routerdc: 基于查询的双对比学习路由器，用于组装大型语言模型. 神经信息处理系统进展, 37: 66305–66328, 2024.
- [9] Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, 和 Kannan Ramchandran. Embedllm: 学习大型语言模型的紧凑表示. arXiv 预印本 arXiv:2410.02223, 2024 年。
- [10] Yi-Kai Zhang, De-Chuan Zhan, 和 Han-Jia Ye. 能力指令微调：动态 LLM 路由的新范式. arXiv 预印本 arXiv:2502.17282, 2025 年。
- [11] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang 和 James Zou. 混合智能体增强大型语言模型能力. arXiv 预印本 arXiv:2406.04692, 2024 年。

- [12] Dawei Li, Zhen Tan, Peijia Qian, Yifan Li, Kumar Satvik Chaudhary, Lijie Hu, and Jiayi Shen. Smoa: Improving multi-agent large language models with sparse mixture-of-agents. *arXiv preprint arXiv:2411.03284*, 2024.
- [13] Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. Rethinking mixture-of-agents: Is mixing different large language models beneficial? *arXiv preprint arXiv:2502.00674*, 2025.
- [14] Justin Chih-Yao Chen, Sukwon Yun, Elias Stengel-Eskin, Tianlong Chen, and Mohit Bansal. Symbolic mixture-of-experts: Adaptive skill-based routing for heterogeneous reasoning, 2025. URL <https://arxiv.org/abs/2503.05641>.
- [15] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.
- [16] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, pages 281–298. University of California press, 1967.
- [17] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.
- [18] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.
- [19] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17:395–416, 2007.
- [20] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114, 1996.
- [21] Jianhao Chen, Zishuo Xun, Bocheng Zhou, Han Qi, Qiaosheng Zhang, Yang Chen, Wei Hu, Yuzhong Qu, Wanli Ouyang, and Shuyue Hu. Do we truly need so many samples? multi-llm repeated sampling efficiently scales test-time compute, 2025. URL <https://arxiv.org/abs/2504.00762>.
- [22] Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *arXiv preprint arXiv:2407.06089*, 2024.
- [23] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.
- [24] Hangfan Zhang, Zhiyao Cui, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. If multi-agent debate is the answer, what is the question? *arXiv preprint arXiv:2502.08788*, 2025.
- [25] Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*, 2025.
- [26] Ziyu Wan, Yunxiang Li, Yan Song, Hanjing Wang, Linyi Yang, Mark Schmidt, Jun Wang, Weinan Zhang, Shuyue Hu, and Ying Wen. Rema: Learning to meta-think for llms with multi-agent reinforcement learning. *arXiv preprint arXiv:2503.09501*, 2025.
- [27] Shenghe Zheng, Hongzhi Wang, Chenyu Huang, Xiaohui Wang, Tao Chen, Jiayuan Fan, Shuyue Hu, and Peng Ye. Decouple and orthogonalize: A data-free framework for lora merging. *arXiv preprint arXiv:2505.15875*, 2025.
- [28] Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2023.

- [12] Dawei Li, Zhen Tan, Peijia Qian, Yifan Li, Kumar Satvik Chaudhary, Lijie Hu 和 Jiayi Shen。Smoa: 利用稀疏混合智能体改进多智能体大语言模型。arXiv 预印本 arXiv:2411.03284, 2024 年。
- [13] 李闻哲, 林勇, 夏梦舟, 金驰。重新思考混合智能体: 混合不同大语言模型是否有益? arXiv 预印本 arXiv:2502.00674, 2025 年。
- [14] Justin Chih-Yao Chen, Sukwon Yun, Elias Stengel-Eskin, Tianlong Chen 和 Mohit Bansal。符号混合专家: 面向异构推理的自适应技能路由, 2025 年。
URL <https://arxiv.org/abs/2503.05641>。
- [15] Stuart Lloyd. 最小二乘量化在 PCM 中的应用。IEEE 信息论汇刊, 28 (2): 129–137, 1982 年。
- [16] James MacQueen. 多元观测分类与分析的一些方法。
载于第五届伯克利数理统计与概率研讨会论文集, 第 1 卷: 统计学, 第 5 卷, 第 281–298 页。加州大学出版社, 1967 年。
- [17] Joe H Ward Jr. 用于优化目标函数的层次分组方法。美国统计协会杂志, 58(301): 236–244, 1963 年。
- [18] Arthur P Dempster, Nan M Laird 和 Donald B Rubin。基于不完全数据的极大似然估计
通过 EM 算法处理数据。皇家统计学会杂志: B 辑 (方法学)》, 39 (1): 1–22, 1977 年。
- [19] Ulrike Von Luxburg. 谱聚类教程。统计与计算》, 17: 395–416。
2007.
- [20] Tian Zhang, Raghu Ramakrishnan 和 Miron Livny。Birch: 一种适用于超大型数据库的高效数据聚类方法。ACM SIGMOD 记录, 25(2): 103–114, 1996 年。
- [21] Jianhao Chen, Zishuo Xun, Bocheng Zhou, Han Qi, Qiaosheng Zhang, Yang Chen, Wei Hu, Yuzhong Qu, Wanli Ouyang 和 Shuyue Hu。我们真的需要这么多样本吗? 多 LLM 重复采样高效扩展测试时计算, 2025 年。网址 <https://arxiv.org/abs/2504.00762>。
- [22] Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia 和 Jiajun Zhang。合并、集成、并合作! 大语言模型时代协作策略综述。arXiv 预印本 arXiv:2407.06089, 2024 年。
- [23] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf 维斯特, 和张祥良。基于大语言模型的多智能体: 进展与挑战综述。arXiv 预印本 arXiv:2402.01680, 2024 年。
- [24] 张航帆, 崔智尧, 王鑫润, 张乔生, 王震, 吴定豪,
和 Shuyue Hu。如果多智能体辩论是答案, 那么问题是什么? arXiv 预印本 arXiv:2502.08788, 2025 年。
- [25] Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, 以及伊戈尔·莫达奇。多智能体微调: 通过多样化推理链实现自我改进。arXiv 预印本 arXiv:2501.05707, 2025 年。
- [26] 万子宇, 李云翔, 宋岩, 王汉晶, 杨林毅, 马克·施密特, 王军,
Weinan Zhang, Shuyue Hu, and Ying Wen. Rema: 通过多智能体强化学习让 LLM 学会元思考。arXiv 预印本 arXiv:2503.09501, 2025 年。
- [27] Shenghe Zheng, Hongzhi Wang, Chenyu Huang, Xiaohui Wang, Tao Chen, Jiayuan Fan, Shuyue 胡, 以及彭烨。解耦与正交化: 一种无数据的 LoRA 合并框架。arXiv 预印本 arXiv:2505.15875, 2025 年。
- [28] 陈灵娇, 马泰·扎哈里亚, 和詹姆斯·邹。FrugalGPT: 如何使用大型语言模型
在降低成本的同时提升性能。机器学习研究汇刊》
2023.

- [29] Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. Large language model routing with benchmark datasets. In *First Conference on Language Modeling*, 2023.
- [30] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In *The Thirteenth International Conference on Learning Representations*, 2024.
- [31] Tao Feng, Yanzhen Shen, and Jiaxuan You. Graphrouter: A graph-based router for LLM selections. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=eU39PDsZtT>.
- [32] Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. Learning to decode collaboratively with multiple language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12974–12990, 2024.
- [33] Zhongzhan Huang, Guoming Ling, Vincent S Liang, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms. *arXiv preprint arXiv:2503.10657*, 2025.
- [34] Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. Routing to the expert: Efficient reward-guided ensemble of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1964–1974, 2024.
- [35] Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. Routerbench: A benchmark for multi-llm routing system. In *Agentic Markets Workshop at ICML 2024*, 2024.
- [36] Wittawat Jitkittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, Zifeng Wang, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, Aditya Krishna Menon, and Sanjiv Kumar. Universal model routing for efficient llm inference. *arXiv preprint arXiv:2502.08773*, 2025.
- [37] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- [38] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- [39] Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, and Kai Chen. Are your llms capable of stable reasoning? *arXiv preprint arXiv:2412.13147*, 2024.
- [40] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- [41] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgén Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

- [29] Tal Shnitzer、Anthony Ou、Mírian Silva、Kate Soule、Yuekai Sun、Justin Solomon、Neil Thompson 儿子，以及 Mikhail Yurochkin。基于基准数据集的大型语言模型路由。见《第一届语言建模会议》，2023 年。
- [30] Isaac Ong、Amjad Almahairi、Vincent Wu、Wei-Lin Chiang、Tianhao Wu、Joseph E Gonzalez、M Waleed Kadous 和 Ion Stoica。Routellm：从偏好数据中学习路由 LLMs。见第十三届国际学习表征会议，2024 年。
- [31] Tao Feng、Yanzhen Shen 和 Jiaxuan You。Graphrouter：一种基于图的 LLM 路由器选择。见第十三届国际学习表征会议，2025 年。
URL <https://openreview.net/forum?id=eU39PDsZtT>。
- [32] Zejiang Shen、Hunter Lang、Bailin Wang、Yoon Kim 和 David Sontag。学习与多个语言模型协作解码。收录于第 62 届年会论文集。计算语言学协会（第 1 卷：长文），第 12974–12990 页，2024 年。
- [33] 黄钟展，凌国明，梁文森，林宇培，陈彦东，陈珊珊钟、何峰武和梁林。Routereval：一个用于路由 LLM 以探索 LLM 模型级扩展的综合基准。arXiv 预印本 arXiv:2503.10657，2025 年。
- [34] 卢克明、袁弘毅、林润基、林俊阳、袁正、周畅和景仁 Zhou。路由至专家：基于奖励引导的大型语言模型高效集成。载于《2024 年北美计算语言学协会会议论文集：人类语言技术（第 1 卷：长论文）》，第 1964–1974 页，2024 年。
- [35] Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer 和 Shriyash Kaustubh Upadhyay。Routerbench：多 LLM 路由系统的基准测试。收录于 ICML 2024 的 Agentic Markets 研讨会，2024 年。
- [36] Wittawat Jitkrittum、Harikrishna Narasimhan、Ankit Singh Rawat、Jeevesh Juneja、Zifeng Wang、Chen-Yu Lee、Pradeep Shenoy、Rina Panigrahy、Aditya Krishna Menon 和 Sanjiv Kumar。用于高效 LLM 推理的通用模型路由。arXiv 预印本 arXiv:2502.08773，2025 年。
- [37] Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。自一致性提升了语言模型中的思维链推理能力。arXiv 预印本 arXiv:2203.11171，2022 年。
- [38] Hunter Lightman、Vineet Kosaraju、Yura Burda、Harri Edwards、Bowen Baker、Teddy Lee、Jan Leike、John Schulman、Ilya Sutskever 和 Karl Cobbe。让我们逐步验证。arXiv 预印本 arXiv:2305.20050，2023 年。
- [39] Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, 和 Kai Chen。你们的 LLMs 是否具备稳定推理能力？arXiv 预印本 arXiv:2412.13147，2024 年。
- [40] Jacob Austin、Augustus Odena、Maxwell Nye、Maarten Bosma、Henryk Michalewski、David Dohan、Ellen Jiang、Carrie Cai、Michael Terry、Quoc Le 等人。利用大型语言模型进行程序合成。arXiv 预印本 arXiv:2108.07732，2021 年。
- [41] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, 和 Wojciech

- [42] Kaijing Ma, Xinrun Du, Yunran Wang, Haoran Zhang, Zhoufutu Wen, Xingwei Qu, Jian Yang, Jiaheng Liu, Minghao Liu, Xiang Yue, Wenhao Huang, and Ge Zhang. Kor-bench: Benchmarking language models on knowledge-orthogonal reasoning tasks, 2024. URL <https://arxiv.org/abs/2410.06526>.
- [43] Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. On memorization of large language models in logical reasoning. *arXiv preprint arXiv:2410.23123*, 2024.
- [44] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.
- [45] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- [46] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.
- [47] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Ti67584b98>.
- [48] Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. Finqa: A dataset of numerical reasoning over financial data. *Proceedings of EMNLP 2021*, 2021.
- [49] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- [50] Mats Byrkjeland, Frederik Gørvell de Lichtenberg, and Björn Gambäck. Ternary twitter sentiment classification with distant supervision and sentiment-specific word embeddings. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 97–106, 2018.
- [51] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2019. URL <https://arxiv.org/abs/1810.02508>.
- [52] Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*, 2024.
- [53] Hannah McLean Babe, Sydney Nguyen, Yangtian Zi, Arjun Guha, Molly Q Feldman, and Carolyn Jane Anderson. Studenteval: A benchmark of student-written prompts for large language models of code. *arXiv preprint arXiv:2306.04556*, 2023.
- [54] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.
- [55] Zahra Rahimi, Mohammad Moein Shirzady, Zeinab Taghavi, and Hossein Sameti. Nimz at semeval-2024 task 9: Evaluating methods in solving brainteasers defying commonsense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 148–154, 2024.

- [42] 马凯静, 杜鑫润, 王云然, 张浩然, 温周福图, 曲星伟, 简杨佳恒、刘明浩、刘明浩、岳翔、黄文浩、张戈。Kor-bench: 在知识正交推理任务上对语言模型进行基准测试, 2024 年。网址: <https://arxiv.org/abs/2410.06526>。
- [43] 谢楚林、黄杨思博、张弛远、余达、陈欣云、林宇辰、李波, Badih Ghazi 和 Ravi Kumar。关于大型语言模型在逻辑推理中的记忆化问题。arXiv 预印本 arXiv:2410.23123, 2024 年。
- [44] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, 和 Jason Wei。挑战 big-bench 任务以及思维链能否解决它们。arXiv 预印本 arXiv:2210.09261, 2022 年。
- [45] Peter Clark、Isaac Cowhey、Oren Etzioni、Tushar Khot、Ashish Sabharwal、Carissa Schoenick, 以及 Oyvind Tafjord。认为你已经解决了问答问题? 试试 ARC, 即 AI2 推理挑战。arXiv 预印本 arXiv:1803.05457, 2018 年。
- [46] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, 等。Mmlu-pro: 一个更稳健且更具挑战性的多任务语言理解基准。arXiv 预印本 arXiv:2406.01574, 2024。
- [47] David Rein、Betty Li Hou、Asa Cooper Stickland、Jackson Petty、Richard Yuanzhe Pang、Julien Dirani、Julian Michael 和 Samuel R. Bowman。GPQA: 一项研究生级别的谷歌 proof q&a 基准测试。收录于第一届语言建模会议, 2024 年。网址: <https://openreview.net/forum?id=Ti67584b98>。
- [48] Zhiyu Chen、Wenhu Chen、Charese Smiley、Sameena Shah、Iana Borova、Dylan Langdon、Reema Moussa、Matt Beane、Ting-Hao Huang、Bryan Routledge 和 William Yang Wang。Finqa: 金融数据数值推理数据集。《MNL2021 论文集》, 2021 年。
- [49] Di Jin、Eileen Pan、Nassim Oufattole、Wei-Hung Weng、Hanyi Fang 和 Peter Szolovits。什么这位患者患有什么疾病? 一个来自医学考试的大规模开放域问答数据集。《应用科学》, 11(14): 6421, 2021 年。
- [50] Mats Byrkjeland、Frederik Gørvell de Lichtenberg 和 Björn Gambäck。三元推特基于远程监督和情感特定词嵌入的情感分类。见《第九届计算主观性、情感与社交媒体分析方法研讨会论文集》, 第 97–106 页, 2018 年。
- [51] Soujanya Poria、Devamanyu Hazarika、Navonil Majumder、Gautam Naik、Erik Cambria 和 Rada Mihalcea。Meld: 用于对话中情绪识别的多模态多方数据集, 2019 年。URL <https://arxiv.org/abs/1810.02508>。
- [52] 刘宏伟, 郑子龙, 乔宇轩, 段浩东, 费志伟, 周凤哲, 温伟张松阳、张大华和林凯。MathBench: 通过分层数学基准评估 LLM 的理论与应用能力。arXiv 预印本 arXiv:2405.12209, 2024 年。
- [53] Hannah McLean Babe、Sydney Nguyen、Yangtian Zi、Arjun Guha、Molly Q Feldman 和 Carolyn Jane Anderson。Studenteval: 面向代码大语言模型的学生编写提示词基准。arXiv 预印本 arXiv:2306.04556, 2023 年。
- [54] Keisuke Sakaguchi、Ronan Le Bras、Chandra Bhagavatula 和 Yejin Choi。Winogrande: 一个在规模上的对抗性维诺格拉德模式挑战。《美国计算机学会通讯》, 64(9): 99–106, 2021 年。
- [55] 扎赫拉·拉希米, 穆罕默德·莫因·希尔扎迪, 泽伊纳布·塔加维, 侯赛因·萨梅蒂。Nimz 在 semeval-2024 任务 9: 评估解决违背常识的脑筋急转弯的方法。见《第十八届国际语义评测研讨会 (SemEval-2024) 论文集》, 第 148–154 页, 2024 年。

- [56] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.
- [57] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.
- [58] OpenAI. New embedding models and api updates. <https://openai.com/index/new-embedding-models-and-api-updates/>, January 2024. Accessed: 2025-05-10.
- [59] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.
- [60] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.
- [61] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.
- [62] Shangbin Feng, Zifeng Wang, Yike Wang, Sayna Ebrahimi, Hamid Palangi, Lesly Miculicich, Achin Kulshrestha, Nathalie Rauschmayr, Yejin Choi, Yulia Tsvetkov, et al. Model swarms: Collaborative search to adapt llm experts via swarm intelligence. *arXiv preprint arXiv:2410.11163*, 2024.
- [63] Chongsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. In *First Conference on Language Modeling*, 2023.
- [64] Yiqun Zhang, Peng Ye, Xiaocui Yang, Shi Feng, Shufei Zhang, Lei Bai, Wanli Ouyang, and Shuyue Hu. Nature-inspired population-based evolution of large language models. *arXiv preprint arXiv:2503.01155*, 2025.
- [65] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- [66] Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. What matters for model merging at scale? *arXiv preprint arXiv:2410.03617*, 2024.
- [67] Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, Chao Li, Sheng Xu, Dezhi Chen, Yun Chen, Zuo Bai, and Liwen Zhang. Fin-r1: A large language model for financial reasoning through reinforcement learning, 2025. URL <https://arxiv.org/abs/2503.16252>.
- [68] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [69] Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-coder technical report, 2024. URL <https://arxiv.org/abs/2409.12186>.
- [70] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

- [56] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, 和 Shuzi Niu. Dailydialog: 一个手动标注的多轮对话数据集. arXiv 预印本 arXiv:1710.03957, 2017.
- [57] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian 和 Zheng Liu. Bge m3-embedding: 通过自知识蒸馏实现的多语言、多功能、多粒度文本嵌入, 2024 年。
- [58] OpenAI. 新的嵌入模型与 API 更新. <https://openai.com/index/new-embedding-models-and-api-updates/>, 2024 年 1 月。访问日期: 2025 年 5 月 10 日。
- [59] 李泽涵, 张鑫, 张彦昭, 龙定坤, 谢鹏军, 张美山。
面向通用文本嵌入的多阶段对比学习. arXiv 预印本, arXiv:2308.03281, 2023 年。
- [60] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel 和 Mohit Bansal. Ties-merging: 解决模型合并时的干扰问题。《神经信息处理系统进展》, 36: 7093–7115, 2023 年。
- [61] 于乐, 于博文, 于海洋, 黄飞, 李永斌. 语言模型是超级 mario: 从同源模型中吸收能力作为免费午餐。见第四十一届国际机器学习大会, 2024 年。
- [62] Shangbin Feng, Zifeng Wang, Yike Wang, Sayna Ebrahimi, Hamid Palangi, Lesly Miculicich, Achin Kulshrestha, Nathalie Rauschmayr, Yejin Choi, Yulia Tsvetkov 等人。模型群体: 通过群体智能协作搜索以适配 LLM 专家. arXiv 预印本 arXiv:2410.11163, 2024 年。
- [63] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du 和 Min Lin. Lorahub: 通过动态 LoRA 组合实现高效的跨任务泛化。载于《语言建模首届会议》, 2023 年。
- [64] 张逸群、叶鹏、杨晓翠、冯时、张书飞、白磊、欧阳万里, 以及 Shuyue Hu. 自然启发的基于种群的演化大型语言模型. arXiv 预印本 arXiv:2503.01155, 2025.
- [65] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu 陈等人. Lora: 大型语言模型的低秩适配。载于国际学习表征会议, 2022 年。
- [66] 普拉蒂克·亚达夫、图·武、乔纳森·赖、亚历山德拉·克罗诺普卢、马纳尔·法鲁基、莫希特·班萨尔, 以及 Tsendsuren Munkhdalai. 大规模模型合并中什么最重要? arXiv 预印本 arXiv:2410.03617, 2024 年。
- [67] 刘兆伟, 郭鑫, 楼方琦, 曾凌峰, 牛金义, 王梓轩, 徐佳杰, 魏格蔡子伟、杨子维、赵学谦、李超、盛旭、陈德智、陈云、陈佐白和张立文. Fin-r1: 通过强化学习实现金融推理的大型语言模型, 2025 年。网址 <https://arxiv.org/abs/2503.16252>。
- [68] 杨安、杨宝松、张北辰、惠彬源、郑波、于博文、陈正元李, Dayiheng Liu, Fei Huang, Haoran Wei, 等. Qwen2.5 技术报告. arXiv 预印本 arXiv:2412.15115, 2024 年。
- [69] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun 张, 博文·余, 克明·陆, 凯·邓, 杨·范, 一昌·张, 安·杨, 瑞·门, 飞·黄, 博·郑, 一波·苗, 尚浩然·权, 云龙·冯, 兴章·任, 宣成·任, 景润·周, 俊阳·林. Qwen2.5-coder 技术报告, 2024 年。URL: <https://arxiv.org/abs/2409.12186>。
- [70] 郭大亚, 杨德健, 张浩伟, 宋俊晓, 张若愚, 徐润欣, 朱启豪, Shirong Ma, Peiyi Wang, Xiao Bi, 等. Deepseek-r1: 通过强化学习激励 LLMs 推理能力. arXiv 预印本 arXiv:2501.12948, 2025.

- [71] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [72] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- [73] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.
- [74] Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL <https://arxiv.org/abs/2503.01743>.
- [75] TII Team. The falcon 3 family of open models, December 2024.
- [76] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- [77] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafford, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024. URL <https://arxiv.org/abs/2501.00656>.
- [78] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaying Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024.

- [71] Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、艾哈迈德·阿尔-达赫勒、艾莎·莱特曼、阿基尔·马图尔、艾伦·谢尔滕、亚历克斯·沃恩等人。《Llama 3 模型系列》。arXiv 预印本, arXiv:2407.21783, 2024 年。
- [72] Gemma 团队、摩根·里维埃、什蕾娅·帕塔克、皮埃尔·朱塞佩·塞萨、卡西迪·哈丁, 苏里亚·布帕蒂拉朱, 莱昂纳尔·胡塞诺, 托马斯·梅斯纳德, 博巴克·沙赫里亚里, 亚历山大·拉梅, 等。Gemma 2: 在实用规模上改进开放语言模型。arXiv 预印本 arXiv:2408.00118, 2024 年。
- [73] GLM 团队, 曾傲寒, 徐斌, 王博文, 张晨辉, 尹达, 张丹, 迭戈·罗哈斯、Guanyu Feng、Hanlin Zhao 等人。Chatglm: 从 glm-130b 到 glm-4 all tools 的大型语言模型系列。arXiv 预印本 arXiv:2406.12793, 2024 年。
- [74] Microsoft、Abdelrahman Abouelenin、Atabak Ashfaq、Adam Atkinson、Hany Awadalla、Nguyen Bach、Jianmin Bao、Alon Benhaim、Martin Cai、Vishrav Chaudhary、Congcong Chen、Dong Chen、Dongdong Chen、Junkun Chen、Weizhu Chen、Yen-Chun Chen、Yiling Chen、Qi Dai、Xiyang Dai、Ruchao Fan、Mei Gao、Min Gao、Amit Garg、Abhishek Goswami、Junheng Hao、Amr Hendy、Yuxuan Hu、Xin Jin、Mahmoud Khademi、Dongwoo Kim、Young Jin Kim、Gina Lee、Jinyu Li、Yunsheng Li、Chen Liang、Xihui Lin、Zeqi Lin、Mengchen Liu、Yang Liu、Gilsinia Lopez、Chong Luo、Piyush Madan、Vadim Mazalov、Arindam Mitra、Ali Mousavi、Anh Nguyen、Jing Pan、Daniel Perez-Becker、Jacob Platin、Thomas Portet、Kai Qiu、Bo Ren、Liliang Ren、Sambuddha Roy、Ning Shang、Yelong Shen、Saksham Singhal、Subhojit Som、Xia Song、Tetyana Sych、Praneetha Vaddamanu、Shuohang Wang、Yiming Wang、Zhenghao Wang、Haibin Wu、Haoran Xu、Weijian Xu、Yifan Yang、Ziyi Yang、Donghan Yu、Ishmam Zabir、Jianwen Zhang、Li Lyna Zhang、Yunan Zhang 和 Xiren Zhou。Phi-4-mini 技术报告: 通过混合 LoRA 微调实现通用推理。arXiv 预印本 arXiv:2409.00511, 2024 年。
- [75] TII 团队。Falcon 3 系列开放模型, 2024 年 12 月。
- [76] Albert Q. Jiang、Alexandre Sablayrolles、Arthur Mensch、Chris Bamford、Devendra Singh Chaplot、Diego de las Casas、Florian Bressand、Gianna Lengyel、Guillaume Lample、Lucile Saulnier、Lélio Renard Lavaud、Marie-Anne Lachaux、Pierre Stock、Teven Le Scao、Thibaut Lavril、Thomas Wang、Timothée Lacroix 和 William El Sayed。Mistral 7b, 2023 年。URL: <https://arxiv.org/abs/2310.06825>
- [77] OLMo 团队, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith 和 Hannaneh Hajishirzi。2 olmo 2 furious。arXiv 预印本 arXiv:2501.00586, 2024 年。URL: <https://arxiv.org/abs/2501.00586>
- [78] 蔡政, 曹茂松, 陈浩炯, 陈凯, 陈科宇, 陈欣, 陈勋, 陈泽辉 Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, 和 Dahua Lin。Internlm2 技术报告。2024 年。

- [79] Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. Hermes 3 technical report, 2024. URL <https://arxiv.org/abs/2408.11857>.
- [80] Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs, 2025. URL <https://arxiv.org/abs/2504.00993>.
- [81] Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzec, Pablo Ribalta, Monika Katariya, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, Eric Chung, and Chris Alexiuk. Llama-nemotron: Efficient reasoning models, 2025. URL <https://arxiv.org/abs/2505.00949>.
- [82] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.
- [83] 01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2025. URL <https://arxiv.org/abs/2403.04652>.

A Technical Appendices and Supplementary Material

A.1 Further Related Work

Mixture-based methods This line of research typically processes each query using multiple models in parallel to generate several candidate responses, and then synthesizes these into a final output using an aggregator model. Mixture-of-Agents (MoA) [11], a notable example, surpasses GPT-4 Omni by leveraging iterative interactions among multi-level agents composed of a pool of medium-sized models (over 70B parameters). Sparse-MOA [12] leverages response selection and early stopping to sparsify information flow, utilizes role-playing for enhanced diversity, achieving comparable performance to standard MoA at a reduced cost. Self-MoA [13] demonstrates that prioritizing

- [79] Ryan Teknium、Jeffrey Quesnelle 和 Chen Guang。Hermes 3 技术报告，2024 年。网址：<https://arxiv.org/abs/2408.11857>。
- [80] Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, 和 Yuyin Zhou. Medreason: 通过知识图谱在 LLMs 中激发事实性医学推理步骤，2025 年。URL <https://arxiv.org/abs/2504.00993>。
- [81] Akhiad Bercovich、Itay Levy、Izik Golan、Mohammad Dabbah、Ran El-Yaniv、Omri Puny、Ido Galil、Zach Moshe、Tomer Ronen、Najeeb Nabwani、Ido Shahaf、Oren Tropp、Ehud Karpas、Ran Zilberstein、Jiaqi Zeng、Soumye Singhal、Alexander Bukharin、Yian Zhang、Tugrul Konuk、Gerald Shen、Ameya Sunil Mahabaleshwarkar、Bilal Kartal、Yoshi Suhara、Olivier Delalleau、Zijia Chen、Zhilin Wang、David Mosallanezhad、Adi Renduchintala、Haifeng Qian、Dima Rekesh、Fei Jia、Somshubra Majumdar、Vahid Noroozi、Wasi Uddin Ahmad、Sean Narenthiran、Aleksander Ficek、Mehrzad Samadi、Jocelyn Huang、Siddhartha Jain、Igor Gitman、Ivan Moshkov、Wei Du、Shubham Toshniwal、George Armstrong、Branislav Kisacanin、Matvei Novikov、Daria Gitman、Evelina Bakhturina、Jane Polak Scowcroft、John Kamalu、Dan Su、Kezhi Kong、Markus Kliegl、Rabeeh Karimi、Ying Lin、Sanjeev Satheesh、Jupinder Parmar、Pritam Gundecha、Brandon Norick、Joseph Jennings、Shrimai Prabhumoye、Syeda Nahida Akter、Mostofa Patwary、Abhinav Khattar、Deepak Narayanan、Roger Waleffe、Jimmy Zhang、Bor-Yiing Su、Guyue Huang、Terry Kong、Parth Chadha、Sahil Jain、Christine Harvey、Elad Segal、Jining Huang、Sergey Kashirsky、Robert McQueen、Izzy Putterman、George Lam、Arun Venkatesan、Sherry Wu、Vinh Nguyen、Manoj Kilaru、Andrew Wang、Anna Warno Abhilash Somasamudramath、Sandip Bhaskar、Maka Dong、Nave Assaf、Shahar Mor、Omer Ullman Argov、Scot Junkin、Oleksandr Romanenko、Pedro Larroy、Monika Katariya、Marco Rovinelli、Viji Balas、Nicholas Edelman、Anahita Bhiwandiwalla、Muthu Subramaniam、Smita Ithape、Karthik Ramamoorthy、Yuting Wu、Suguna Varshini Velury、Omri Almog、Joyjit Daw、Denys Fridman、Erick Galinkin、Michael Evans、Katherine Luna、Leon Derczynski、Nikki Pope、Eileen Long、Seth Schneider、Guillermo Siman、Tomasz Grzegorzec、Pablo Ribalta、Monika Katariya、Joey URL <https://arxiv.org/abs/2505.00949>。
- [82] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, 和 Hannaneh Hajishirzi. Tulu 3: 推动开放语言模型后训练的前沿. arXiv 预印本 arXiv:2411.15124, 2024.
- [83] 01.AI: Alex Young、陈贝、李超、黄承恩、张戈、管伟、郭银王、李恒、蒋承柱、陈建群、张静、余凯东、刘鹏、刘强、Shawn Yue、杨森彬、杨世明、谢文、黄文浩、胡晓辉、任晓毅、牛欣瑶、彭成晔、李彦鹏、徐玉驰、刘玉东、王悦、蔡宇轩、顾振宇、刘志远、戴宗宏。《01.AI 的开源基础模型》，2025 年。网址：<https://arxiv.org/abs/2403.04652>。

A 技术附录与补充材料

A.1 更多相关工作

基于混合的方法 这一研究方向通常使用多个模型并行处理每个查询，以生成若干候选回复，然后通过一个聚合模型将这些回复综合为最终输出。混合智能体 (Mixture-of-Agents, MoA) [11] 是一个显著例子，它通过利用由一组中等规模模型（参数超过 70B）构成的多层级智能体之间的迭代交互，超越了 GPT-4 Omni。Sparse-MOA [12] 利用回复选择和提前停止来稀疏化信息流，并通过角色扮演增强多样性，在降低成本的同时实现了与标准 MoA 相当的性能。Self-MoA [13] 表明，优先考虑

high-quality outputs from the model is crucial, and shows that selecting the single best-performing model on the current task to serve simultaneously as both Aggregator and Proposer can surpass the standard MoA. Symbolic-MoE [14] introduces a skill-based adaptive routing strategy that leverages an additional SLM to generate skill profiles for each query and candidate SLM, dynamically recruiting the most suitable expert models at the query level.

Compared to router-based approaches, mixture-based methods are training-free and often demonstrate greater performance gains over single models. However, such frameworks heavily depend on hand-crafted prompts and meticulously selected architectural components, such as the careful selection of aggregator models. Moreover, they often entail intricate hybrid (serial-parallel) execution flows, leading to substantial inference overhead. Last but not least, they necessitate the abilities of instruction following and prompt robustness—which are key shortcomings of small language models. Although the *Avengers* does *not* fall within this paradigm, we include comparisons with representative mixture-based methods to highlight our framework’s effectiveness, especially in the context of small language models.

Merging-based methods This line of research combines the parameters of several fine-tuned models (typically derived from the same base model) to produce a single, more versatile model with enhanced task-specific performance and broader generalization. Relevant works include TIES-Merging [60], DARE [61], Model Swarms [62], LoRAHub [63], and GENOME [64]. Notably, recent studies predominantly focus on lightweight parameter merging techniques such as LoRA [65], with limited exploration of full-parameter merging strategies, thereby constraining the performance ceiling. Furthermore, most existing methods assume architectural homogeneity [66], i.e., homogeneous models derived from the same base model. This assumption is fundamentally different from ours. Thus, in this work, we do not compare our framework to this line of research.

A.2 Models

Table 7 shows the 10 mainly models used in the main experiments in our paper.

Table 7: Detailed information of the models, including their abbreviations.

Model	Abbreviation	Parameters (B)
Fin-R1 [67]	Fin-R1	7.61
Qwen2.5-7B-Instruct [68]	Qwen-it	7.61
Qwen2.5-Coder-7B-Instruct [69]	Qwen-Coder	7.61
DeepSeek-R1-Distill-Qwen-7B [70]	DS-Qwen	7.61
cogito-v1-preview-llama-8B	cogito-v1	8.03
Llama-3.1-8B-Instruct [71]	Llama-3.1-it	8.03
Llama-3.1-8B-UltraMedical [3]	UltraMedical	8.03
Granite-3.1-8B-Instruct	Granite-3.1-it	8.03
gemma-2-9b-it [72]	gemma-2-it	9.24
glm-4-9b-chat [73]	glm-4-chat	9.40
Average Parameters	–	8.12

Table 8 presents the remaining 12 models utilized in the experiments detailed in Section 4.3. Notably, we begin our experiments with an initial pool consisting of all 22 models. Subsequently, an automated model selection method is applied to select the 10 models listed in Table 7. All results report in the main text, including those for baseline comparisons, consistently use the same set of models selected by the *Avengers*.

A.3 Datasets

We provide the detailed information of datasets in Table 9.

A.4 Baseline

LLM Router We use Qwen2.5-7B-Instruct as the router, which takes the current query along with descriptions (profiles) of available models to automatically select one or multiple suitable models for

Self-MoA [13] 表明，优先考虑模型的高质量输出至关重要，并展示了在当前任务上选择性能最佳的单一模型同时充当聚合器和提议者，可以超越标准 MoA。Symbolic-MoE [14] 引入了一种基于技能的适应性路由策略，利用额外的 SLM 为每个查询和候选 SLM 生成技能档案，在查询级别动态招募最合适的专家模型。

与基于路由器的方法相比，基于混合的方法无需训练，且通常比单一模型展现出更大的性能提升。然而，此类框架高度依赖手工设计的提示词和精心挑选的架构组件，例如对聚合模型的谨慎选择。此外，它们往往涉及复杂的混合（串行-并行）执行流程，导致显著的推理开销。最后但同样重要的是，它们需要指令遵循和提示鲁棒性的能力——而这正是小型语言模型的关键短板。尽管 Avengers 不属于这一范式，我们仍纳入了与代表性基于混合方法的对比，以突出我们框架的有效性，尤其是在小型语言模型的背景下。

基于合并的方法 这一研究方向将多个微调模型（通常源自同一基础模型）的参数合并，以生成一个更通用、任务性能更强且泛化能力更广的单一模型。相关工作包括 TIES-Merging [60]、DARE [61]、Model Swarms [62]、LoRAHub [63]和 GENOME [64]。值得注意的是，近期研究主要集中于轻量级参数合并技术（如 LoRA [65]），而对全参数合并策略的探索有限，从而限制了性能上限。此外，大多数现有方法假设架构同质性 [66]，即模型源自同一基础模型且结构相同。这一假设与我们的方法存在根本性差异。

因此，在本工作中，我们不将我们的框架与这一研究方向进行比较。

A.2 模型

表 7 展示了我们论文中主要实验所使用的 10 个主要模型。

表 7：模型的详细信息，包括其缩写。

模型	缩写	参数 (B)
Fin-R1 [67]	Fin-R1	7.61
Qwen2.5-7B-Instruct [68]	Qwen-it	7.61
Qwen2.5-Coder-7B-Instruct [69]	Qwen-Coder	7.61
DeepSeek-R1-Distill-Qwen-7B [70]	DS-Qwen	7.61
cogito-v1-preview-llama-8B	cogito-v1	8.03
Llama-3.1-8B-Instruct [71]	Llama-3.1-it	8.03
Llama-3.1-8B-UltraMedical [3]	UltraMedical	8.03
Granite-3.1-8B-Instruct	Granite-3.1-it	8.03
gemma-2-9b-it [72]	gemma-2-it	9.24
glm-4-9b-chat [73]	glm-4-chat	9.40
平均参数	-	8.12

表 8 展示了第 4.3 节详细实验中使用的其余 12 个模型。值得注意的是，我们的实验初始模型池包含全部 22 个模型。随后，采用自动化模型选择方法从其中选出表 7 所列的 10 个模型。正文中报告的所有结果，包括基线对比结果，均一致使用由 Avengers 选出的同一组模型。

A.3 数据集

我们在表 9 中提供了数据集的详细信息。

A.4 基线

LLM 路由器 我们使用 Qwen2.5-7B-Instruct 作为路由器，它接收当前查询以及可用模型的描述（配置文件），自动选择一个或多个合适的模型来

Table 8: Other models.

Model	Parameters (B)
Phi-4-mini-instruct [74]	3.8
Falcon3-7B-Instruct [75]	7
mistral-7b-instruct-v0.3 [76]	7.25
OLMo-2-1124-7B-Instruct [77]	7.3
Eurus-2-7B-PRIME [4]	7.61
Qwen2.5-7B-MATH-Instruct [68]	7.61
internlm3-8b-instruct [78]	8
Hermes-3-Llama-3.1-8B [79]	8.03
Medreason-8B [80]	8.03
Llama-3.1-Nemotron-Nano-8B-v1 [81]	8.03
Llama-3.1-Tulu-3.1-8B [82]	8.03
Yi-1.5-9B-Chat [83]	8.83
Average Parameters	7.46

Table 9: Detailed information of the datasets.

Dataset	Category	Metrics	Size
AIME	Mathematics	Accuracy, 0-shot	60
MATH500 [38]	Mathematics	Accuracy, 0-shot	500
LiveMathBench [39]	Mathematics	Accuracy, 0-shot	140
MBPP [40]	Code	Pass@1, 0-shot	974
HumanEval [41]	Code	Pass@1, 0-shot	164
KORBench [42]	Logic	Accuracy, 3-shot	1 250
Knights and Knaves [43]	Logic	Accuracy, 0-shot	700
BBH [44]	Logic	Accuracy, 3-shot	1 080
ARC Challenge [45]	Knowledge	Accuracy, 0-shot	1 172
MMLUPro [46]	Knowledge	Accuracy, 0-shot	1 000
GPQA [47]	Knowledge	Accuracy, 0-shot	448
FinQA [48]	Knowledge	Accuracy, 0-shot	1 147
MedQA [49]	Knowledge	Accuracy, 0-shot	1 273
EmoryNLP [50]	Affective	Accuracy, 0-shot	697
MELD [51]	Affective	Accuracy, 0-shot	1 232
<i>Out of Distribution Datasets</i>			
MathBench [52]	Mathematics	Accuracy, 0-shot	150
StudentEval [53]	Code	Pass@1, 0-shot	953
Winogrande [54]	Logic	Accuracy, 0-shot	1 267
BRAINTEASER [55]	Knowledge	Accuracy, 0-shot	1 021
DailyDialog [56]	Affective	Accuracy, 0-shot	1 000

the query. This router requires no additional training; it only relies on natural-language model profiles prepared for each model. In our implementation, we introduce an automatic method to generate these model profiles. Specifically, we first collect the performance scores of all models on the validation set. Then, using task descriptions and these scores, we prompt Qwen2.5-7B-Instruct to automatically generate the corresponding model profiles. To ensure consistent performance, we keep the request parameters fixed at temperature = 0.2 and top_p = 1.0 for both profile generation and inference in the LLM router.

RouterDC We employ the official implementation [8], substituting the encoder with gte-qwen2-7B-instruct to ensure a fair comparison with our method. We utilize DeepSpeed for distributed multi-GPU training on eight NVIDIA A800-80G GPUs, setting the batch size per GPU to 8 while keeping all other hyperparameters unchanged. Our implementation comprises approximately 7B trainable parameters. The accuracy curves for both training and test are shown in Figure 5.

MODEL-SAT Since the official implementation provided by [10] is incomplete, we re-implement the primary method, reproduce the code, and report the results. We employ gte-qwen2-7B-instruct as the embedding model and Qwen2.5-7B-Instruct as the language model, connecting them via a

表 8：其他模型。

模型	参数量 (B)
Phi-4-mini-instruct [74]	3.8
Falcon3-7B-Instruct [75]	7
mistral-7b-instruct-v0.3 [76]	7.25
OLMo-2-1124-7B-Instruct [77]	7.3
Eurus-2-7B-PRIME [4]	7.61
Qwen2.5-7B-MATH-Instruct [68]	7.61
internlm3-8b-instruct [78]	8
Hermes-3-Llama-3.1-8B [79]	8.03
Medreason-8B [80]	8.03
Llama-3.1-Nemotron-Nano-8B-v1 [81]	8.03
Llama-3.1-Tulu-3.1-8B [82]	8.03
Yi-1.5-9B-Chat [83]	8.83
平均参数	7.46

表 9：数据集的详细信息。

数据集	类别指标	规模
AIME	数学准确率, 0-shot	60
数学准确率	0-shot	380
HumanEval	代码 Pass@1, 0 样本	500
LiveMathBench [39]	数学 准确率, 0 样本	140
MBPP [40]	代码 Pass@1, 0 样本	974
HumanEval [41]	代码 Pass@1, 0 样本	164
KORBench [42]	逻辑 准确率, 3 样本	1 250
Knights and Knaves [43]	逻辑 准确率, 0 样本	700
BBH [44]	逻辑 准确率, 3 样本	1 080
ARC Challenge [45]	知识 准确率, 0 样本	1 172
MMLUPro [46]	知识 准确率, 0 样本	1 000
GPQA [47]	知识 准确率, 0 样本	448
FinQA [48]	知识 准确率, 0 样本	1 147
MedQA [49]	知识 准确率, 0 样本	1 273
EmoryNLP [50]	情感 准确率, 0 样本	697
MELD [51]	情感 准确率, 0 样本	1 232
分布外数据集		
MathBench [52]	数学 准确率, 0-shot	150
StudentEval [53]	代码 Pass@1, 0-shot	953
Winogrande [54]	逻辑 准确率, 0-shot	1 267
BRAINTEASER [55]	知识 准确率, 0-shot	1 021
DailyDialog [56]	情感 准确率, 0-shot	1 000

该路由器接收当前查询以及可用模型的描述（配置文件），自动为查询选择一个或多个合适的模型。该路由器无需额外训练，仅依赖为每个模型准备的自然语言模型配置文件。在我们的实现中，我们引入了一种自动生成这些模型配置文件的方法。具体而言，我们首先收集所有模型在验证集上的性能得分。然后，利用任务描述和这些得分，我们提示 Qwen2.5-7B-Instruct 自动生成相应的模型配置文件。为确保性能一致性，我们在配置文件生成和 LLM 路由器推理过程中均将请求参数固定为 temperature = 0.2 和 top_p = 1.0。

RouterDC 我们采用官方实现[8]，将编码器替换为 gte-qwen27B-instruct，以确保与我们的方法进行公平比较。我们使用 DeepSpeed 在八块 NVIDIA A800-80G GPU 上进行分布式多 GPU 训练，将每块 GPU 的批大小设置为 8，同时保持所有其他超参数不变。我们的实现包含约 7B 可训练参数。训练和测试的准确率曲线如图 5 所示。

MODEL-SAT 由于[10]提供的官方实现不完整，我们重新实现了主要方法，复现了代码，并报告了结果。我们采用 gte-qwen2-7B-instruct 作为嵌入模型，Qwen2.5-7B-Instruct 作为语言模型，通过

two-layer MLP projector. For efficient training, we leverage DeepSpeed for distributed multi-GPU training across eight NVIDIA A800-80GB GPUs, with a batch size of 2 per GPU. The learning rate is configured at $2e-6$ for both the embedding and language models, while the projector uses a higher learning rate of $1e-4$. Initially, we exclusively fine-tune the projector for approximately 10,000 steps. Following this phase, we continue fine-tuning all model parameters for the remainder of the training process. A warmup ratio of 0.1 is applied to stabilize the early training stages. Our implementation comprises approximately 14B trainable parameters. Figure 6 illustrates the training and test accuracy curves.

EmbedLLM We employ the official implementation [9] and generate query embeddings using gte-qwen2-7B-instruct to ensure a fair comparison with our method, adjusting input layer dimensions accordingly. We increase the batch size to 32,768 for more stable training, keeping all other hyperparameters unchanged. Our implementation comprises approximately 12 million trainable parameters. The training and test accuracy curves are presented in Figure 7.

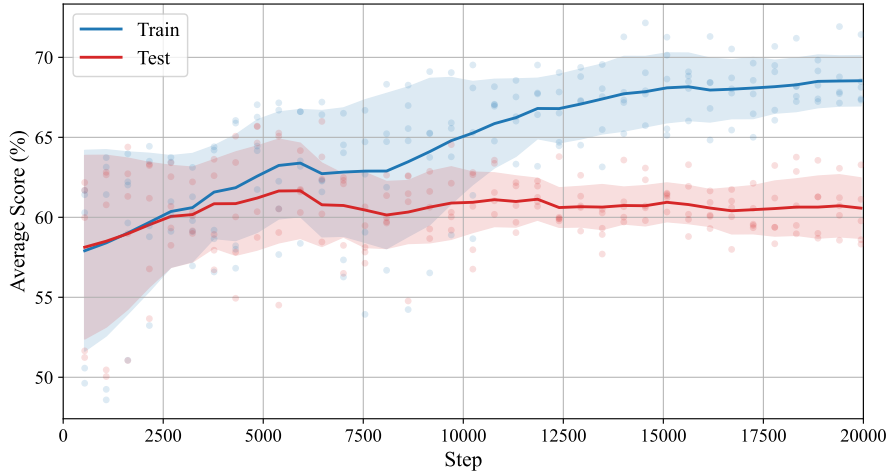


Figure 5: The training and test accuracy curves on RouterDC.

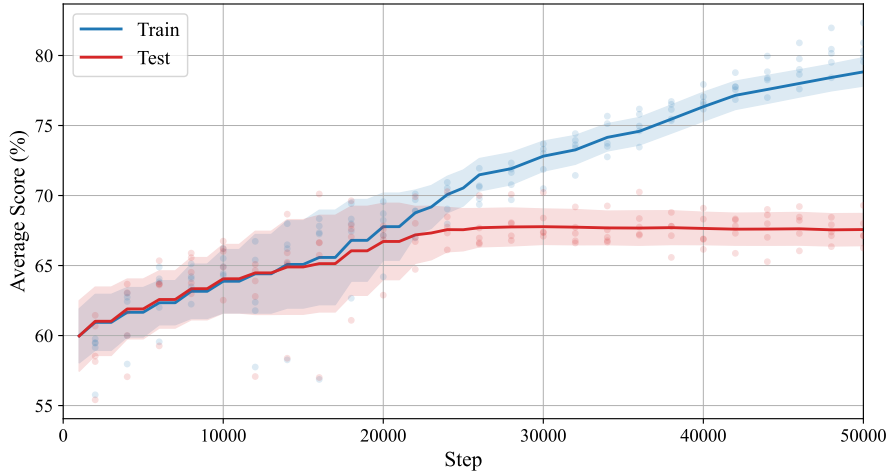


Figure 6: The training and test accuracy curves on MODEL-SAT.

Mixture of Agents We adopt the classic three-layer architecture from the official implementation [11], incorporating all ten candidate models (i.e., each proposer layer includes all ten models, with an aggregator layer summarizing the information). We choose Qwen2.5-7B-Instruct as the aggregator due to its strong performance, long context window, and robust instruction-following

通过一个两层 MLP 投影仪将它们连接起来。为了高效训练，我们利用 DeepSpeed 在八块 NVIDIA A800-80GB GPU 上进行分布式多 GPU 训练，每块 GPU 的批大小为 2。嵌入模型和语言模型的学习率均设置为 $2e-6$ ，而投影仪则使用更高的学习率 $1e-4$ 。最初，我们仅对投影仪进行约 10,000 步的微调。在此阶段之后，我们继续对训练过程剩余部分的所有模型参数进行微调。采用 0.1 的预热比例以稳定早期训练阶段。我们的实现包含约 140 亿个可训练参数。图 6 展示了训练和测试准确率曲线。EmbedLLM 我们采用官方实现[9]，并使用 gte-qwen2-7B-instruct 生成查询嵌入，以确保与我们的方法进行公平比较，并相应调整输入层维度。我们将批量大小增加到 32,768 以实现更稳定的训练，其余超参数保持不变。我们的实现包含约 1200 万个可训练参数。

训练和测试准确率曲线如图 7 所示。

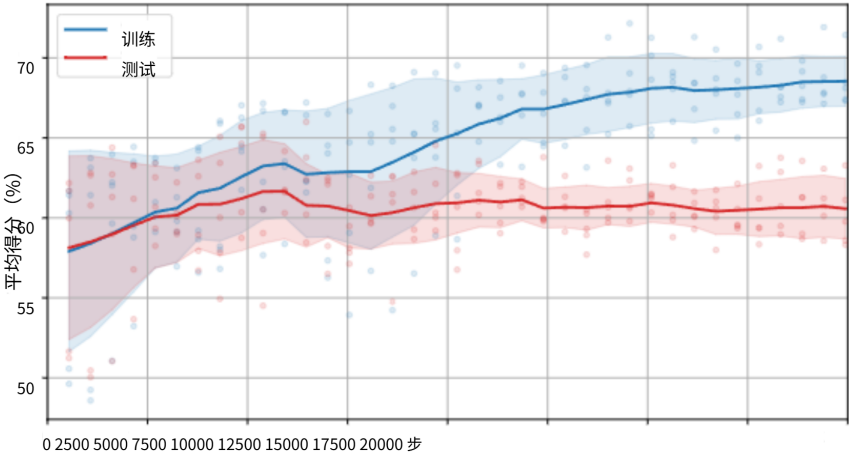


图 5：RouterDC 上的训练与测试准确率曲线。

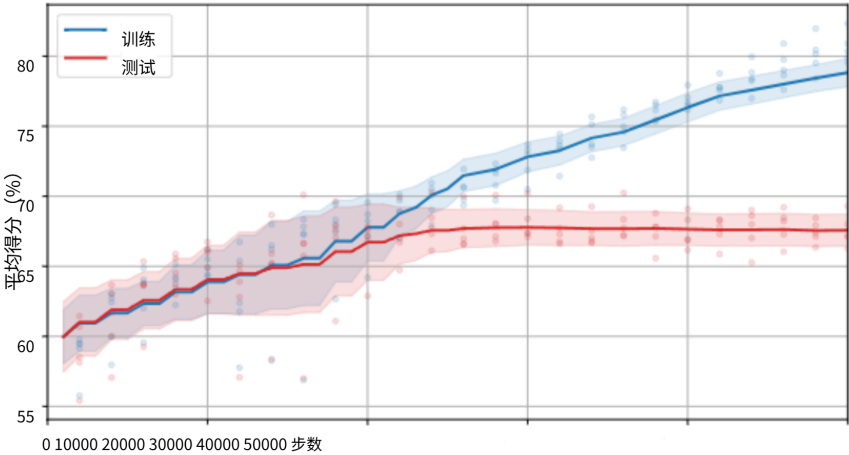


图 6：MODEL-SAT 上的训练和测试准确率曲线。

混合智能体我们采用了官方实现[11]中的经典三层架构，整合了全部十个候选模型（即每个提议层包含所有十个模型，由聚合层汇总信息）。我们选择 Qwen2.5-7B-Instruct 作为聚合器，因其性能优异、上下文窗口长且指令遵循能力强。

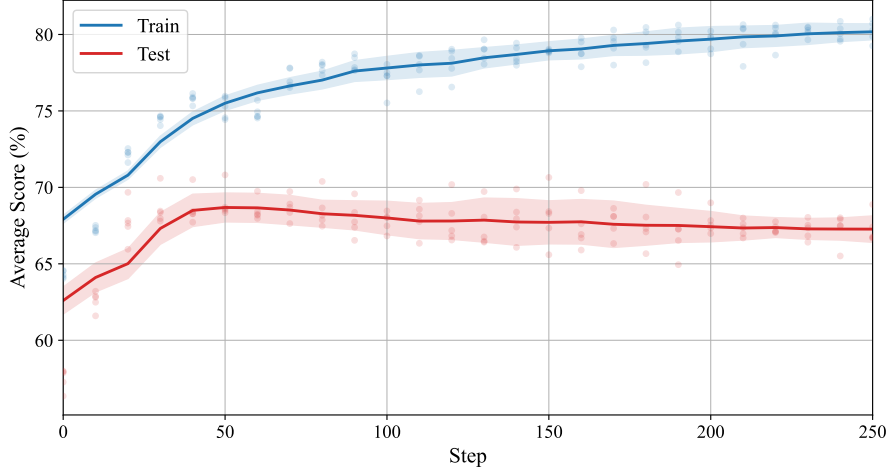


Figure 7: The training and test accuracy curves on EmbedLLM.

capabilities. To ensure that the context length stays within model limits, we set a context window of 6,000 tokens for each proposer layer and 20,000 tokens for the aggregator layer.

Mixture of Agents (Oracle) For a more comprehensive comparison, we enhance the MoA algorithm by adjusting the selection of models in each proposer layer. Specifically, rather than using all ten candidate models at each layer (to avoid performance degradation due to excessively long context windows), we manually select the three best-performing models on the validation set for each specific task as proposers. Nevertheless, Qwen2.5-7B-Instruct remains the aggregator across all tasks. For instance, for the AIME task, we select DeepSeek-Distill-R1-Qwen-7B and Qwen2.5-7B-Instruct as proposers, while for the MELD task, we choose Qwen2.5-7B-Coder-Instruct, Qwen2.5-7B-Instruct, and Fin-R1. The three-layer architecture and context window sizes remained consistent with the original MoA configuration.

A.5 Symbolic Mixture-of-Experts

We employ the official implementation [14], likewise utilizing Qwen2.5-7B-Instruct to generate skill keywords for each model and query. We adopt bge-m3 to generate embeddings for each keyword, thereby achieving superior performance compared to the Sentence-BERT used in the official implementation, while keeping all other hyperparameters unchanged.

B Implementation Details

All 22 LLMs are served on NVIDIA A800-80G GPUs. To speed up inference, every model—including the embedding model—is deployed with vLLM 0.8.4 on the same hardware. Direct (CoT) set temperature = 0.2, top_p = 1.0; voting strategies (Model-Switch, Self-Consistency) set temperature = 0.7, top_p = 1.0 (all other parameters remain at their default values). When using the aggregation strategy, we adopt the same parameters as in Direct (CoT). For reproduce our results, we open-source our code on GitHub: <https://anonymous.4open.science/r/Avengers-2C0D>

B.1 Benchmark Dataset Descriptions

AIME Olympiad-level math benchmark of *American Invitational Mathematics Examination* problems (2024 & 2025); numeric-answer tasks covering algebra, geometry, number theory and combinatorics that push high-school contest reasoning.

MATH500 Subset of Hendrycks MATH: 500 diverse competition problems spanning probability, algebra, geometry and more, sampled for quick yet rigorous evaluation of multi-step symbolic reasoning.

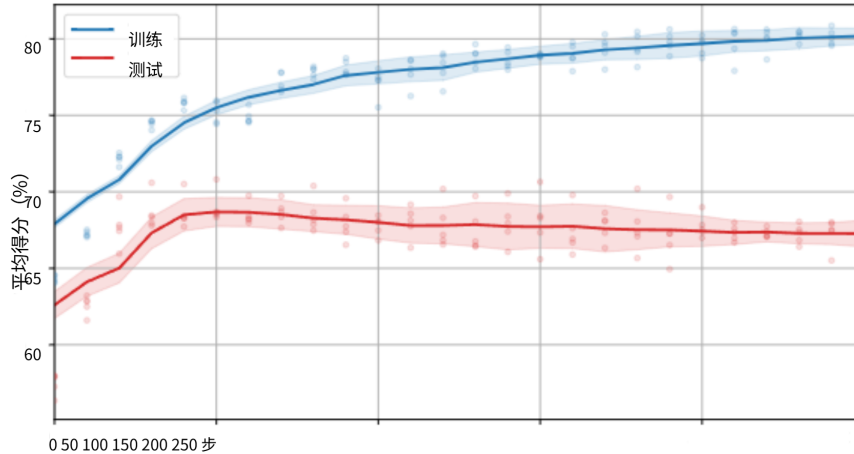


图 7: EmbedLLM 上的训练和测试准确率曲线。

并具备强大的指令遵循能力。为确保上下文长度不超过模型限制，我们为每个提议层设置了 6,000 个令牌的上下文窗口，并为聚合层设置了 20,000 个令牌的上下文窗口。

混合代理 (Oracle) 为了进行更全面的比较，我们通过调整每个提议层中模型的选择来增强 MoA 算法。具体而言，我们并非在每个层使用全部十个候选模型（以避免因上下文窗口过长而导致性能下降），而是手动为每个特定任务选择验证集上表现最佳的三个模型作为提议者。尽管如此，Qwen2.5-7B-Instruct 在所有任务中仍作为聚合器。例如，对于 AIME 任务，我们选择 DeepSeek-Distill-R1-Qwen-7B 和 Qwen2.5-7B-Instruct 作为提议者，而对于 MELD 任务，我们选择 Qwen2.5-7B-Coder-Instruct、Qwen2.5-7B-Instruct 和 Fin-R1。三层架构和上下文窗口大小与原始 MoA 配置保持一致。

A.5 符号混合专家

我们采用官方实现[14]，同样利用 Qwen2.5-7B-Instruct 为每个模型和查询生成技能关键词。我们采用 bge-m3 为每个关键词生成嵌入，从而在性能上优于官方实现中使用的 Sentence-BERT，同时保持所有其他超参数不变。

B 实现细节

全部 22 个 LLM 均部署在 NVIDIA A800-80G GPU 上。为加速推理，所有模型（包括嵌入模型）均使用 vLLM 0.8.4 部署在同一硬件上。直接 (CoT) 策略设置 temperature = 0.2, top_p = 1.0；投票策略 (Model-Switch、Self-Consistency) 设置 temperature = 0.7, top_p = 1.0（其余参数保持默认值）。使用聚合策略时，我们采用与直接 (CoT) 相同的参数。为便于复现我们的结果，我们已在 GitHub 上开源代码：<https://anonymous.4open.science/r/Avengers-2C0D>

B.1 基准数据集描述

AIME 美国数学邀请赛奥林匹克级数学基准测试题

2024 年和 2025 年的问题；涵盖代数、几何、数论和组合数学的数值答案任务，这些任务考验高中竞赛推理能力。

MATH500 是 Hendrycks MATH 的一个子集：包含 500 道涵盖概率、代数等领域的多样化竞赛题目。代数、几何等更多内容，采样用于快速而严格的多步符号推理评估。

LiveMathBench 2024 benchmark with 140 recent questions from CNMO, CCEE, AMC and Putnam competitions, curated to avoid training contamination and probe stable mathematical reasoning in LLMs.

MBPP Mostly Basic Python Programming: $\approx 1\,000$ crowd-sourced entry-level Python tasks, each with docstring, reference solution and three unit tests; used to measure code-generation functional correctness.

HumanEval OpenAI HumanEval: 164 handwritten Python programming problems with hidden unit tests, specifically drafted to eliminate training leakage and evaluated via pass@k.

KORBench Knowledge-Orthogonal Reasoning Benchmark evaluates intrinsic reasoning across five rule-driven categories (Operation, Logic, Cipher, Puzzle, Counterfactual) while minimizing domain-knowledge cues.

Knight and Knaves Logic-puzzle dataset containing 700 scenarios with 2–8 inhabitants where knights always tell the truth and knaves always lie; the task is to deduce each role from their statements.

BBH Big-Bench Hard: 23 of the most challenging BIG-Bench tasks where earlier models lagged human baselines, covering diverse domains such as causal reasoning, word sorting and formal fallacies.

ARCC ARC Challenge—1172 multiple-choice grade-school science questions that simple retrieval baselines miss; assesses open-domain reasoning with a 14 M-sentence supporting corpus

MMLUPro MMLU-Pro extends MMLU with 1000 harder questions across 14 disciplines plus chain-of-thought annotations, offering a more robust university-level multi-task understanding benchmark.

GPQA Graduate-level Physics Qualifying Q&A benchmark designed to be Google-proof; covers classical mechanics, EM, quantum, thermo and more, forcing deep conceptual reasoning.

FinQA Financial numerical-reasoning dataset with 1147 Q&A pairs over 2.8k corporate reports, combining tabular and textual evidence and requiring multi-step arithmetic reasoning.

MedQA Medical QA benchmark built from USMLE and other board exams; 1273 English multiple-choice questions (plus Chinese variants) spanning basic science, diagnosis and treatment.

EmoryNLP Emotion-in-conversation corpus from 97 Friends TV-show episodes, totalling 12 606 utterances labelled with seven discrete emotions for dialogue emotion recognition research.

MELD Multimodal EmotionLines Dataset: 1.4 k multi-speaker dialogues and 13 k utterances from Friends, providing synchronized text, audio and vision plus emotion and sentiment labels.

B.2 Model Descriptions

Fin-R1 Specialises in quantitative reading of corporate data and long-form fiscal reasoning, while maintaining competitive scores on competition-style mathematics and factual QA. Its attention to ledger-like details comes at the cost of brittle performance on paradox-style logic games, underscoring a finance-oriented optimisation.

Qwen2.5-7B-Instruct Pairs top-tier coding accuracy with strong symbolic-math reasoning and achieves the most even spread across general-knowledge and dialogue tasks among its size class. Its moderate showing on intricate logic puzzles hints at a balanced—but not specialised—architecture that trades depth in any single niche for consistently high all-round competence.

Qwen2.5-Coder-7B-Instruct Excels at structured programming challenges and keeps pace with larger peers on algorithmic reasoning, while retaining solid performance in numeric problem-solving and business analytics. It concedes ground on tricky logical riddles and specialised domain questions, revealing a profile tuned first for code reliability rather than broad expert knowledge.

DeepSeek-R1-Distill-Qwen-7B Dominates Olympiad-level mathematics and logic riddles, while retaining competitive breadth across finance and encyclopaedic knowledge tests—a rare feat for its parameter count. The main gap is in highly specialised clinical material and affective-dialogue sensing, pointing to a math-centric curriculum.

LiveMathBench 2024 基准测试，包含来自 CNMO、CCEE、AMC 和 Put-的 140 道最新题目 nam 竞赛，经过精心策划以避免训练数据污染，并测试 LLMs 在稳定数学推理方面的能力。

MBPP（基础 Python 编程）：约 1000 个众包的入门级 Python 任务，每个任务包含文档字符串、参考解决方案和三个单元测试；用于衡量代码生成的功能正确性。

HumanEval OpenAI HumanEval：164 个手写 Python 编程问题，附带隐藏单元测试，专门设计以避免训练数据泄漏，并通过 pass@k 进行评估。

KORBench 知识正交推理基准通过五个

规则驱动类别（运算、逻辑、密码、谜题、反事实）评估内在推理能力，同时最小化领域知识线索。

Knight and Knaves 逻辑谜题数据集，包含 700 个场景，每个场景有 2 至 8 名居民，其中骑士总是说真话，而骗子总是说谎；任务是根据他们的陈述推断出每个人的角色。

BBH Big-Bench Hard：23 项最具挑战性的 BIG-Bench 任务，早期模型在这些任务上表现落后人类基线，涵盖因果推理、单词排序和形式谬误等多个领域。

ARCC ARC 挑战——1172 道多项选择的小学科学问题，简单的检索基线无法解答；通过包含 1400 万句子的支持语料库评估开放域推理能力 MMLUPro MMLU-Pro 在 MMLU 基础上扩展了 1000 道更难的题目，涵盖 14 个学科，并增加了链式思维链标注，提供了一个更稳健的大学水平多任务理解基准。

GPQA 研究生水平物理资格考试问答基准，设计为“谷歌不可解”；涵盖经典力学、电磁学、量子力学、热力学等，迫使进行深层概念推理。

FinQA 金融数值推理数据集，包含 1147 个问答对，覆盖 2.8 千份企业报告，结合表格与文本证据，需要多步算术推理。

MedQA 医学问答基准，基于 USMLE 及其他委员会考试构建；包含 1273 道英文选择题（另有中文变体），涵盖基础科学、诊断与治疗。

埃默里 NLP 对话情感语料库，源自 97 集《老友记》电视剧，共计 12,606 条带有七种离散情感标注的话语，用于对话情感识别研究。

MELD 多模态情感对话数据集：来自《老友记》的 1.4 千个多说话人对话和 1.3 万条话语，提供同步的文本、音频和视觉信息，以及情感和情绪标签。

B.2 模型描述

Fin-R1 专注于企业数据的定量解读和长篇幅财务推理，同时在竞赛式数学和事实问答上保持有竞争力的得分。其对账本式细节的关注，以在悖论式逻辑游戏上表现脆弱为代价，凸显了以金融为导向的优化。

Qwen2.5-7B-Instruct 将顶级的编码准确性与强大的符号数学推理能力相结合，在其同尺寸类别中，在通用知识和对话任务上实现了最均衡的表现。它在复杂逻辑谜题上的中等表现暗示了一种均衡但非专门化的架构，这种架构以牺牲任何单一领域的深度为代价，换取了持续的高水平全面能力。

Qwen2.5-Coder-7B-Instruct 在结构化编程挑战中表现出色，并在算法推理方面与规模更大的同类模型保持同步，同时在数值问题求解和商业分析方面保持了扎实的性能。它在棘手的逻辑谜题和专业领域问题上有所让步，展现出一种首先为代码可靠性而非广泛专家知识而调优的模型特征。

DeepSeek-R1-Distill-Qwen-7B 在奥林匹克级别的数学和逻辑谜题中占据主导地位，同时在金融和百科知识测试中保持了具有竞争力的广度——对于其参数规模而言，这是一项罕见的成就。主要差距在于高度专业化的临床材料和情感对话感知，这表明其训练课程以数学为

Llama-3.1-8B-Instruct Balances solid software-engineering skills with reliable performance in factual and medical queries, and even handles tricky reasoning suites respectably. The trade-off is weaker abstract-math fluency and lower success on self-contradictory dialogue puzzles, placing it as a versatile generalist with coding leanings.

Granite-3.1-8B-Instruct Shows surprising strength on science-fact retrieval and spreadsheet-style arithmetic despite muted results on code and higher maths. The model is clearly tuned for concise knowledge lookup and tabular reasoning, but its conservative logical toolkit becomes evident on riddle-like tasks.

Llama-3.1-8B-UltraMedical Unrivalled within the cohort on clinical examination questions and therapeutic reasoning, confirming a medical specialist orientation. Outside its home turf it posts fair results on general reading but lags behind peers on advanced logic and algorithmic coding, illustrating the classic depth-versus-breadth trade-off.

cogito-v1-preview-llama-8B Stands out on open-ended reasoning and multi-domain science questions, matching larger models on financial and medical queries despite modest maths prowess. Slightly weaker numerical intuition is offset by a marked advantage on complex instruction-following and argumentative tasks, making it a “thinker” rather than a “calculator”.

gemma-2-9b-it Offers dependable code generation and strong comprehension of technical prose, coupled with above-average domain knowledge in healthcare and finance. Elementary competition maths and deception-based puzzles remain its softer spots, signalling a language-focused training bias over abstract symbol manipulation.

glm-4-9b-chat Delivers well-rounded scores, peaking on encyclopaedic comprehension and practical coding, yet rarely leading except in fact-heavy domains. Its numbers reveal a steady, middle-of-the-pack stance—above smaller baselines in most categories, but outshone by maths specialists and domain experts on their respective turf.

Parameter Setting of Cluster Methods We implement all clustering algorithms using `scikit-learn` with the following default hyperparameters:

- **K-Means** [15, 16]: We set the number of clusters $K = 64$, with the “k-means++” initialization method, automatic initialization count (`n_init=‘auto’`), a maximum of 1000 iterations, and the “elkan” algorithm.
- **Hierarchical Clustering** [17] (Hier-Clustering): We use the Euclidean distance metric with the “ward” linkage method and set the number of clusters to $K = 64$.
- **Gaussian Mixture Models** [18] (GMM): We specify 64 components, use full covariance matrices, and set the number of initializations (`n_init`) to 10. The maximum iterations are set to 100, with verbosity level 2, printing intermediate results every 10 iterations.
- **Balanced Iterative Reducing and Clustering using Hierarchies** [20] (BIRCH): We use the default threshold of 0.5, a branching factor of 50, and set the number of clusters to $K = 64$.
- **Spectral Clustering** (Spec-Clustering): We set the number of clusters to $K = 64$, use the RBF affinity matrix with the parameter `gamma` set to 0.1.

B.3 Further Analysis

Test Size We further investigate how the proportion of the test set impacts the performance of the *Avengers*. Specifically, we vary the test set proportion from 0.05 to 0.95 and analyze its effect on model performance. As shown in Figure 8, results indicate that the *Avengers* achieves the best overall performance when the test set proportion is approximately between 0.3 and 0.35. Additionally, it is noteworthy that with only 30% of the data used to fit the clustering model, the *Avengers* already surpasses the performance of GPT-4.1, demonstrating the high efficiency of our method in terms of labeled data utilization.

Further Analysis of Ensemble Strategy In this paragraph, we further analysis the impact of ensemble strategy. As shown in Table 10, model-Switch achieves competitive overall performance, notably excelling on knowledge tasks by combining predictions from two models for complementary strengths. Among these, Self-Consistency emerges as the best strategy due to its simple yet effective

Llama-3.1-8B-Instruct 在扎实的软件工程技能与可靠的性能之间取得了平衡

事实性和医学类查询表现出色，甚至在棘手的推理套件上也有不错的表现。其代价是抽象数学流畅度较弱，在自相矛盾的对话谜题上成功率较低，使其成为一款偏向编码能力的多面手

Granite-3.1-8B-Instruct 在科学事实检索和电子表格类任务上展现出惊人的实力。

在代码和高等数学方面表现平平，但在算术上却有所建树。该模型显然针对简洁的知识查询和表格推理进行了调优，但在谜题类任务上，其保守的逻辑工具集便显露无遗。

Llama-3.1-8B-UltraMedical 在临床考试题目上于该队列中无出其右，且

治疗性推理，确认了医学专科方向。在其专业领域之外，它在通用阅读方面表现尚可，但在高级逻辑和算法编码上落后于同类模型，体现了经典的深度与广度权衡。

cogito-v1-preview-llama-8B 在开放式推理和多领域科学问题上表现突出。

在金融和医疗查询方面可与更大模型匹敌，尽管数学能力平平。稍弱的数值直觉被在复杂指令遵循和论证性任务上的显著优势所抵消，使其成为“思考者”而非“计算器”。

gemma-2-9b-it 提供可靠的代码生成能力，并对技术性文本有很强理解力，

在医疗和金融领域拥有高于平均水平的领域知识。基础竞赛数学和基于欺骗的谜题仍是其薄弱环节，表明其训练偏向语言理解而非抽象符号操作。

glm-4-9b-chat 在各项得分上表现均衡，在百科全书式理解和实用

编码能力尚可，但除事实密集型领域外鲜有领先。其数据展现出一种稳定的中游姿态——在大多数类别中优于较小的基线模型，但在各自擅长的领域被数学专家和领域专家超越。

聚类方法的参数设置 我们使用 scikit-learn 实现所有聚类算法，并采用以下默认超参数：

- K-Means [15, 16]：我们设置聚类数量 $K = 64$ ，采用“k-means++”初始化方法，自动初始化次数 ($n_init = \text{“auto”}$)，最大迭代次数为 1000，并使用“elkan”算法。
- 层次聚类 [17] (Hier-Clustering)：我们使用欧氏距离度量，采用“ward”连接方法，并将聚类数量设置为 $K = 64$ 。
- 高斯混合模型 [18] (GMM)：我们指定 64 个分量，使用全协方差矩阵，并将初始化次数 (n_init) 设置为 10。最大迭代次数设为 100，详细程度级别为 2，每 10 次迭代打印一次中间结果。
- 使用层次结构的平衡迭代归约与聚类 [20] (BIRCH)：我们使用默认阈值 0.5，分支因子为 50，并将聚类数设置为 $K = 64$ 。
- 谱聚类 (Spec-Clustering)：我们将聚类数量设置为 $K = 64$ ，使用 RBF 亲和矩阵，并将参数 γ 设置为 0.1。

B.3 进一步分析

测试集大小 我们进一步研究了测试集比例对复仇者联盟性能的影响。具体而言，我们将测试集比例从 0.05 变化到 0.95，并分析其对模型性能的影响。如图 8 所示，结果表明，当测试集比例大约在 0.3 到 0.35 之间时，复仇者联盟取得了最佳的整体性能。此外，值得注意的是，仅使用 30% 的数据来拟合聚类模型，复仇者联盟就已经超越了 GPT-4.1 的性能，这证明了我们的方法在标注数据利用方面的高效率。

集成策略的进一步分析 在本段中，我们进一步分析了集成策略的影响。如表 10 所示，模型切换策略取得了具有竞争力的整体性能，尤其在知识任务上表现出色，它通过结合两个模型的预测来实现优势互补。其中，自一致性策略因其简单而有效的特点，成为最佳策略。

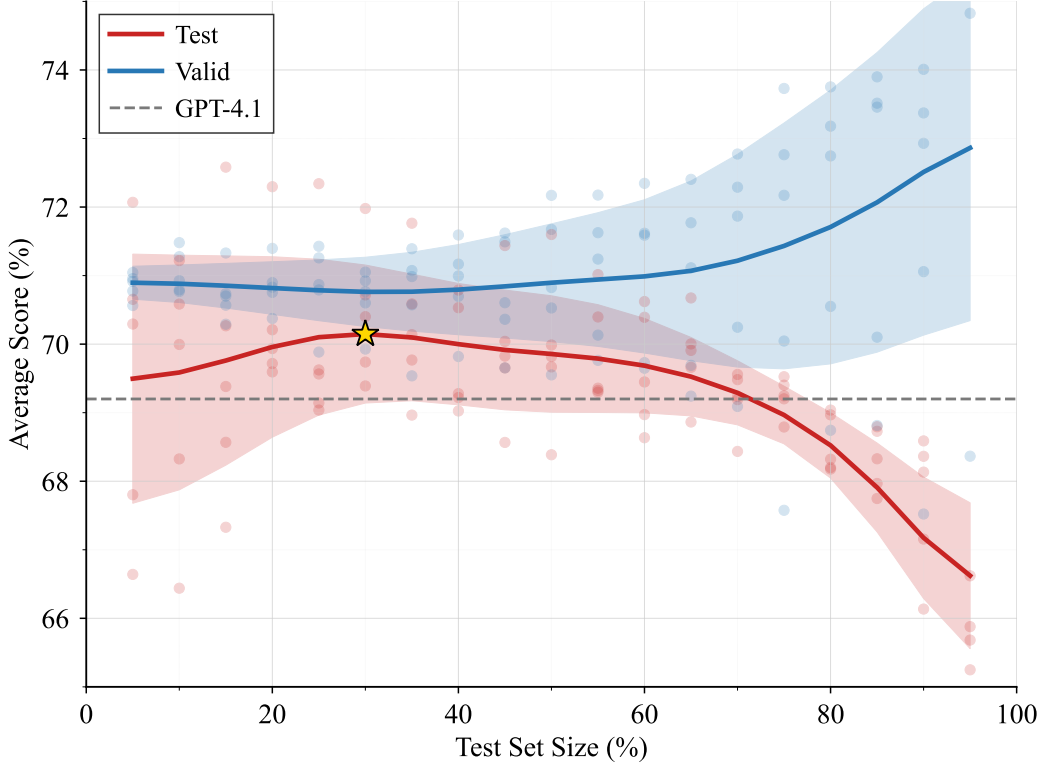


Figure 8: The impact of test set proportion on the *Avengers* performance.

procedure of majority voting, reducing single-sample inference errors and maintaining robustness across tasks.

Table 10: Ablation on the *Avengers* ensemble strategies. Numbers are the average accuracy for four categories (Mathematics, Logic, Knowledge, Affective); the Code category is omitted because there is no straightforward way to compute a self-consistency score for program generation.

Ensemble Strategy	Mathematics	Logic	Knowledge	Affective	Avg.
Direct (CoT)	66.89	60.77	65.84	43.27	59.19
Aggregation	59.39	53.89	68.91	44.75	56.73
Model-Switch	71.11	71.27	72.22	51.09	66.42
Self-Consistency	73.42	72.06	71.96	<u>50.50</u>	66.98

B.4 Speculative Insights on the Effectiveness of Clustering-Based Routing

In this paragraph, we would like to answer why our upscaling is effective without requiring training. Overall, Figure 4 illustrates three distinct regimes: **(i) Under-clustering** ($K < 14$): As the cluster number K initially increases from very small values, both training and testing performance rapidly improve. **(ii) Stability** ($14 < K < 140$): As K grows further to a suitable intermediate range, both validation and test performance stabilize and remain closely matched. During this phase, the validation set performance improves gradually while the test set performance remains stable and consistently surpasses GPT-4.1. Notably, the cluster number ($K = 64$) chosen for all our main experiments lies in the middle of this region, validating the robustness and stability of our parameter selection. **(iii) Over-clustering** ($K > 140$): With a further increase in cluster number, the validation set performance gradually approaches the Oracle level, as each query nearly forms its own cluster, thus precisely finding the optimal model for each query. However, the performance on the test set sharply deteriorates, exhibiting typical over-clustering behavior.

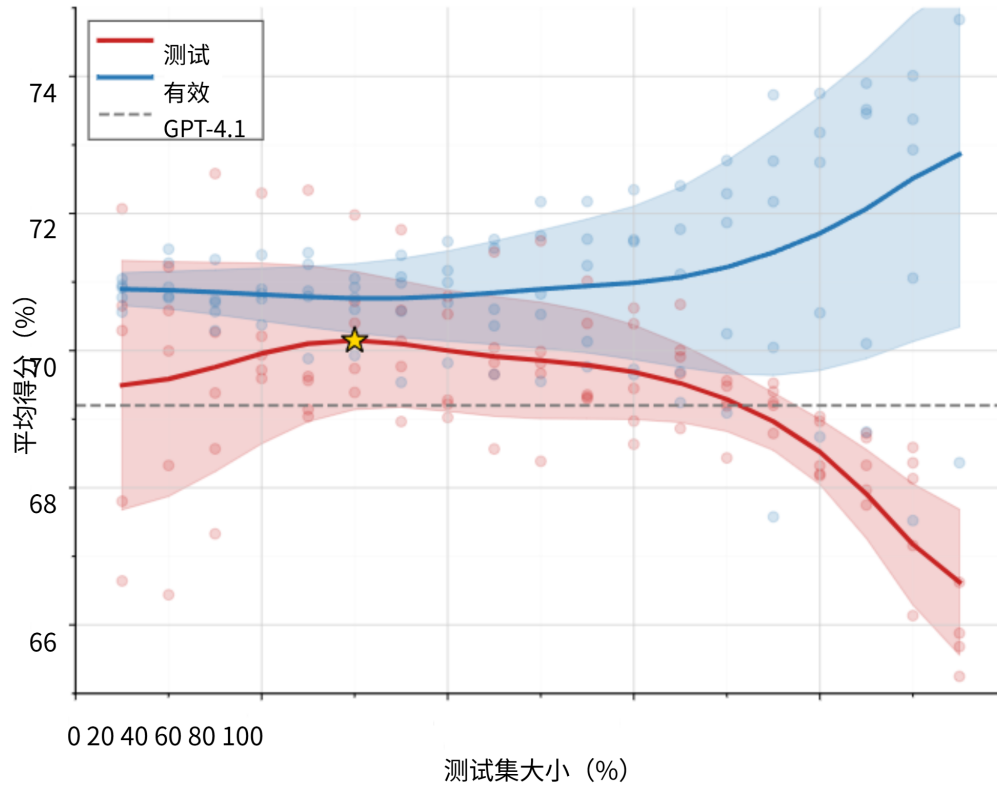


图 8：测试集比例对复仇者联盟性能的影响。

自治性因其简单而有效的多数投票机制而成为最佳策略，能够减少单样本推理误差，并在各任务中保持稳健性。

表 10：关于“复仇者联盟”集成策略的消融实验。数字为四个类别（数学、逻辑、知识、情感）的平均准确率；代码类别因无法直接计算程序生成的自我一致性分数而被省略。

集成策略	数学	逻辑	知识	情感	平均
直接 (CoT)	66.89	60.77	65.84	43.27	59.19
聚合	59.39	53.89	68.91	44.75	56.73
模型切换	71.11	71.27	72.22	51.09	66.42
自治性	73.42	72.06	71.96	50.50	66.98

B.4 关于基于聚类的路由有效性的推测性见解

在本段中，我们希望回答为什么我们的上采样方法无需训练即可有效。总体而言，图 4 展示了三种不同的区域：(i) 欠聚类 ($K < 14$)：随着聚类数 K 从极小值开始增加，训练和测试性能均迅速提升。(ii) 稳定区 ($14 < K < 140$)：随着 K 进一步增大至合适的中间范围，验证集和测试集性能趋于稳定并保持高度一致。在此阶段，验证集性能逐步提升，而测试集性能保持稳定并持续超越 GPT-4.1。值得注意的是，我们所有主要实验所选的聚类数 ($K = 64$) 恰好位于该区域的中间位置，验证了参数选择的稳健性和稳定性。(iii) 过聚类 ($K > 140$)：随着聚类数进一步增加，验证集性能逐渐接近 Oracle 水平，因为每个查询几乎形成独立的聚类，从而为每个查询精确找到最优模型。然而，测试集性能急剧下降，表现出典型的过聚类行为。

A clear understanding of this “sweet-spot vs. over-clustering” pattern also offers a lens for interpreting why heavily trained router-based methods tend to struggle in our setting. We hypothesize that other router-based methods (RouterDC, EmbedLLM and MODEL-SAT) requiring additional training are prone to quickly entering the over-clustering (overfitting) region illustrated in Figure 4, due to the introduction of extra learnable parameters and complex network structures (such as embedding models, MLP layers, or even LLMs). These methods typically optimize directly toward maximizing validation-set (training) performance, and combined with their large number of parameters, this significantly accelerates overfitting and consequently harms their generalization performance. Moreover, identifying a simple yet effective early-stopping criterion to prevent overfitting becomes challenging in practice. This explains why, in Tables 2 and 3, we report peak performance on the test set rather than the validation set for these trained methods. In the Appendix A.4, we further provide training curves for these router-based methods, which exhibit a similar overfitting pattern to that observed in Figure 4, further validating our hypothesis.

This experiment thus highlights the efficiency and simplicity of the *Avengers*, which effectively balances the precision of model selection and generalization capability through clustering without introducing any learnable parameters. This design avoids the overfitting pitfalls common to more complex methods.

B.5 Model Usage Distribution within the *Avengers*

Figure 9 provides a detailed breakdown of the model usage distribution within the *Avengers* framework across various benchmark datasets. The analysis reveals that the *Avengers* demonstrates a clear capability for automatically identifying models that balance strong performance with complementary strengths. While some well-known general-purpose models, such as Qwen2.5-7B-Instruct, dominate usage across multiple tasks, less prominent models also play critical roles for specific tasks. For instance, Fin-R1—originally developed for financial-domain tasks—is heavily utilized in the ARC Challenge (~50% of queries), while LLaMA-3.1-8B-UltraMedical, developed by a small academic team, is notably selected for approximately 40% of queries in the MedQA dataset. These results underscore not only the versatility and effectiveness of the *Avengers* approach but also highlight the significant yet often overlooked potential of specialized, community-contributed open-source models, contributing towards a more diverse and inclusive AI ecosystem.

B.6 Limitation

Similar to other router-based methods, the *Avengers* requires setting aside a portion of the data as a validation set to obtain capability profiles.

Although we conduct experiments across 15 datasets spanning five task categories and performed additional out-of-distribution evaluations on five extra datasets, the *Avengers* has not yet been validated on chat-based tasks involving cross-cultural, multilingual, or conversational contexts. This limitation arises primarily due to the subjective nature of these tasks, which lack standardized objective metrics, making their performance challenging to measure reliably.

Currently, our experiments with the *Avengers* focus exclusively on integrating the capabilities of multiple small-scale (approximately 7B) language models, and we have not yet explored its applicability to smaller-scale models. The choice of 7B-scale models is primarily motivated by the abundant availability and active community support for open-source models of this scale.

对“最佳平衡点与过度聚类”模式的清晰理解，也为解读为何重度训练的路由器方法在我们的设置中往往表现不佳提供了视角。我们假设，其他需要额外训练的路由器方法（RouterDC、EmbedLLM 和 MODEL-SAT）由于引入了额外的可学习参数和复杂网络结构（如嵌入模型、MLP 层，甚至 LLMs），容易迅速进入图 4 所示的过度聚类（过拟合）区域。这些方法通常直接以最大化验证集（训练）性能为优化目标，加之其参数数量庞大，显著加速了过拟合，从而损害了其泛化性能。此外，在实践中，要找到一个简单而有效的早停标准来防止过拟合变得颇具挑战。这解释了为何在表 2 和表 3 中，我们针对这些训练方法报告的是测试集上的峰值性能，而非验证集上的。在附录 A 中，我们进一步提供了这些基于路由方法的训练曲线，其表现出与图 4 中观察到的类似过拟合模式，进一步验证了我们的假设。

本次实验因此凸显了 Avengers 的高效性与简洁性，它通过聚类在模型选择的精确性与泛化能力之间实现了有效平衡，且未引入任何可学习参数。这一设计避免了更复杂方法中常见的过拟合陷阱。

B.5 Avengers 内部的模型使用分布

图 9 详细展示了 Avengers 框架在不同基准测试数据集中的模型使用分布情况。分析表明，Avengers 展现出清晰的能力，能够自动识别在性能强大与优势互补之间取得平衡的模型。虽然一些知名的通用模型（如 Qwen2.5-7B-Instruct）在多个任务中占据主导地位，但一些不太知名的模型在特定任务中也发挥着关键作用。例如，Fin-R1——最初为金融领域任务而开发——在 ARC Challenge 中被大量使用（约占查询量的 50%），而由小型学术团队开发的 LLaMA-3.1-8B-UltraMedical 在 MedQA 数据集中被显著选中，约占查询量的 40%。这些结果不仅凸显了 Avengers 方法的通用性和有效性，还强调了专业化、社区贡献的开源模型所具有的重大但往往被忽视的潜力，为构建更加多元化和包容性的 AI 生态系统做出了贡献。

B.6 局限性

与其他基于路由的方法类似，Avengers 需要预留一部分数据作为验证集，以获取能力画像。

尽管我们在涵盖五个任务类别的 15 个数据集上进行了实验，并在另外五个数据集上进行了额外的分布外评估，但 Avengers 尚未在涉及跨文化、多语言或对话情境的聊天型任务上得到验证。这一局限性主要源于这些任务的主观性，缺乏标准化的客观指标，使得其性能难以可靠衡量。

目前，我们对 Avengers 的实验仅专注于整合多个小规模（约 7B）语言模型的能力，尚未探索其对更小规模模型的适用性。选择 7B 规模的模型主要基于该规模开源模型的丰富可用性和活跃的社区支持。

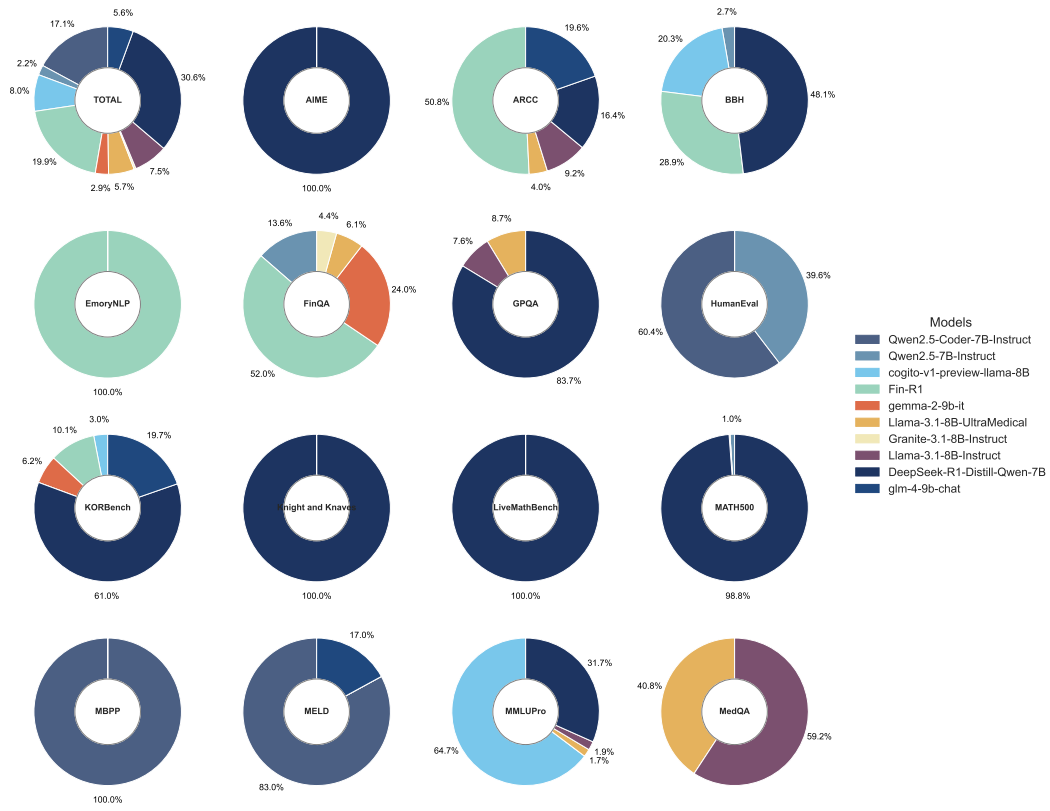


Figure 9: Distribution of the routing results of the *Avengers* across tasks.

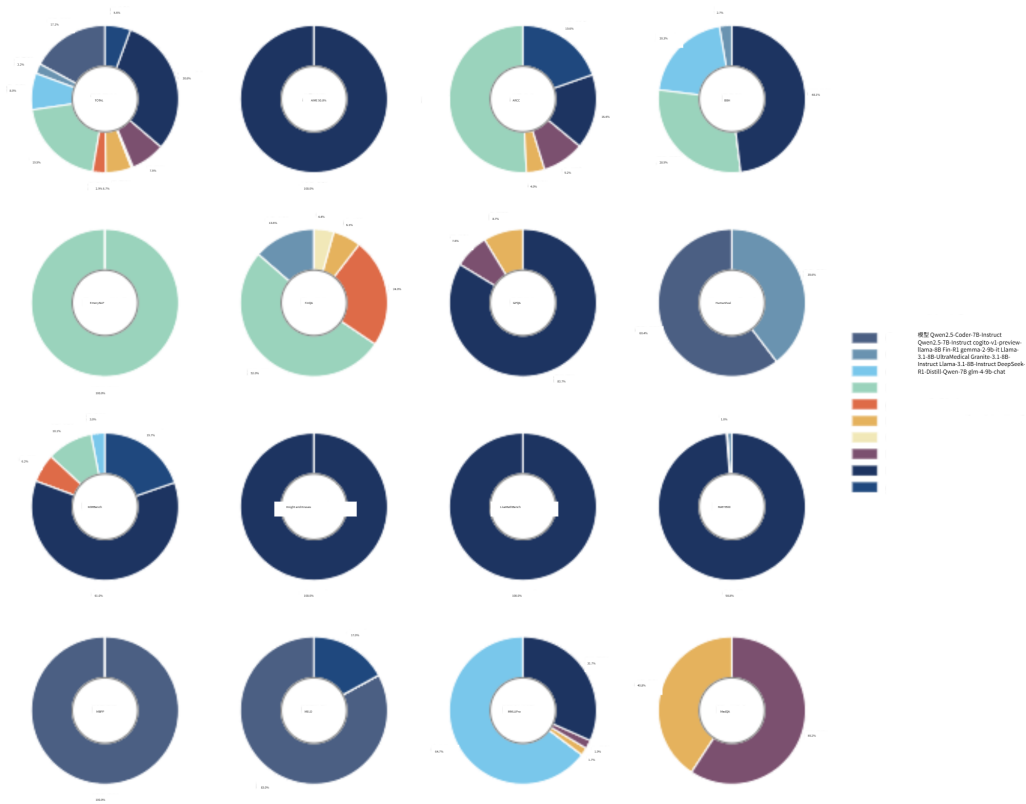


图 9：复仇者联盟在各任务中的路由结果分布。