Generalizing Test-time Compute-optimal Scaling as an Optimizable Graph

Fali Wang[1]*, **Jihai Chen**[1]*, **Shuhua Yang**[1], **Runxue Bao**[2], **Tianxiang Zhao**[1], **Zhiwei Zhang**[1], **Xianfeng Tang**[3], **Hui Liu**[3], **Qi He**[3], **Suhang Wang**[1]
[1]The Pennsylvania State University, University Park, PA, USA
[2]University of Pittsburgh, Pittsburgh, PA, USA
[3]Microsoft, Mountain View, CA, USA
{fqw5095,jjc8069,szw494}@psu.edu

Abstract

Test-Time Scaling (TTS) improves large language models (LLMs) by allocating additional computation during inference, typically through parallel, sequential, or hybrid scaling. However, prior studies often assume fixed collaboration architectures (e.g., topologies) and single-model usage, overlooking that optimal architectures and model combinations can vary across tasks. Therefore, we study the novel problem of *searching for compute-optimal model combinations and architectures in TTS under a fixed budget.* We formalize it as a multi-LLM collaboration graph, where nodes encode roles and LLM model assignments, and edges capture information flow. This problem is challenging because (i) the combinatorial search space is prohibitively large, and (ii) task-specific requirements demand tailored designs. To this end, we formulate the problem as probabilistic graph optimization and, via pilot experiments, derive three empirical insights into TTS collaboration graphs. Guided by these insights, we propose Agent-REINFORCE, an LLM-agent-augmented framework that mirrors the REINFORCE pipeline by mapping *sampling–gradient–update* to *sampling–feedback–update*, where feedback serves as a textual gradient to update the probabilistic graph and efficiently search for optimal multi-LLM collaboration graphs. Experiments show that Agent-REINFORCE outperforms both traditional and LLM-based baselines in sample efficiency and search performance, and effectively identifies optimal graphs under joint objectives of accuracy and inference latency. Our code is publicly available at link.

1 Introduction

Test-time scaling (TTS) aims to enhance large language models (LLMs) by allocating additional computational resources during inference [5, 49]. Prior studies have primarily investigated two architectures: (i) *parallel scaling* [55, 5], which samples multiple outputs independently to increase solution diversity and aggregates them, making it suitable for tasks with uncertain or diverse solution paths; and (ii) *sequential scaling* [37, 49], which iteratively refines a single output and is well-suited for tasks that require step-by-step reasoning (see Fig. 2 (a)(b)). Fusing the two, hybrid architectures have also been proposed, using predefined hybrid structures to combine the advantages of both [4, 49] (see Fig. 2 (c)). Despite their effectiveness, we identify two key limitations of existing TTS architectures. **First, TTS architectures are typically predefined and static, with fixed topologies across tasks.** However, our analysis shows that different tasks exhibit distinct preferences for architectural patterns, e.g., MATH favors hybrid structures, while MMLU performs better with pure parallel ones (Fig. 1(a)(c)). This suggests that architectures should adapt to task demands.

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

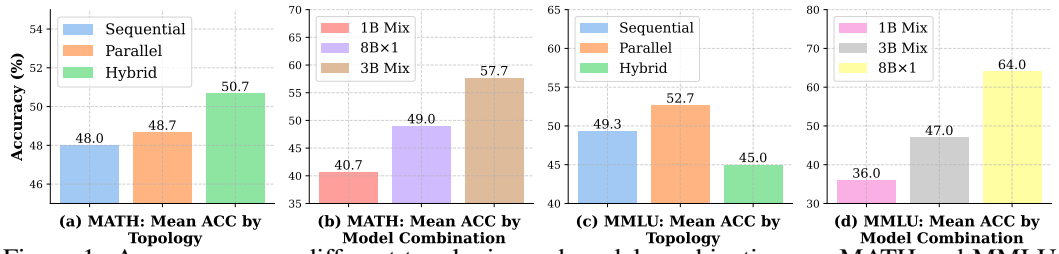

Figure 1: Accuracy across different topologies and model combinations on MATH and MMLU. LLaMA-3 models are used by default. Detailed data is in Appendix A.3.

Second, existing TTS methods usually employ a single LLM for all inference steps. In contrast, multi-LLM ensembles are preferable to leverage heterogeneous LLM skills across tasks [27, 54]. Preliminary results show that MATH benefits from mixtures of 1B–3B, whereas MMLU favors a single 8B (Fig. 1(b)(d)), underscoring the need for adaptive model selection. Overall, *test-time compute-optimal scaling* aims to maximize performance within the inference budget [58], but these findings reveal that *adaptive TTS architectures and model combinations are fundamental challenges for existing methods.*

Motivated by these observations, we study a novel problem: **searching for the compute-optimal architecture and model combination in test-time scaling for a given task**. Formally, given a task, a set of models, and a compute budget, the goal is to find the best configuration that jointly determines architecture and model assignment. Leveraging the inherent graph structure of TTS, we formulate dynamic test-time scaling as constructing a *multi-LLM collaboration graph*, where nodes represent the chosen LLM model with assigned roles (*fuser* for parallel aggregation, *assistant* for sequential refinement), and edges denote information flow. A terminal node aggregates outputs into the final answer (see Fig. 2(d)). This graph view offers a systematic foundation for dynamic optimization. However, two major challenges arise: **(i) The search space is large due to the combinatorial choices of models and topologies, and grows rapidly with the budget.** For example, with 12 nodes, the number of possible graphs ranges from 10^{18} to 10^{26} depending on model diversity (derivation in Appendix A.4). Since evaluating each candidate requires costly inference, brute-force search is infeasible. **(ii) Tailored design requires linking task requirements to optimal TTS search patterns, which relies on an understanding of TTS behaviors.** Prior work shows that performance does not grow monotonically with used budget, implying that optimal allocations are often below the maximum. These insights are key to guiding task-specific searches toward compute-optimal collaboration graphs. To address them, we conducted pilot experiments on TTS behavior analysis, which yielded three empirical insights: (1) Effective collaboration exhibits clear preferences for specific model combinations: tasks favor replication of the strongest model family, and ensembles of small models are preferred when incremental gains are substantial; (2) Both width and depth have task-dependent optima; beyond these points, extra computation will yield negative returns; (3) Graph width and depths are interdependent: growth in one dimension shifts the optimal point of the other.

We operationalize these insights by formulating the search as a probabilistic optimization problem: Learning a distribution over collaboration graphs that jointly determines edges, roles, and model

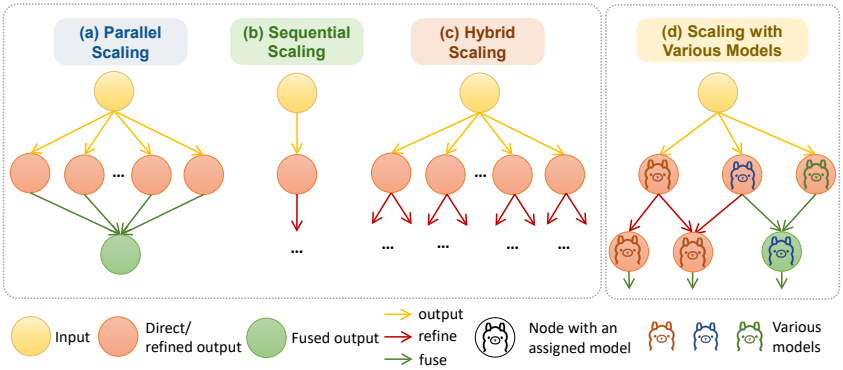

Figure 2: Test-Time Scaling Paradigms: (a–c) Fixed topologies with single-model assignments, and (d) dynamic scaling with diverse models.

assignments under a fixed budget to maximize task-specific performance. The REINFORCE algorithm [57], a gradient-based optimization method, addresses this via a *sample–gradient–update* pipeline that iteratively samples candidates, computes gradients, updates the distribution, and repeats. However, it risks local optima and its inability to incorporate empirical insights. Recent work [33, 70] shows that LLM-based agents are effective planners for hyperparameter optimization, with the unique advantage of leveraging external knowledge. Building on these, we propose **Agent-REINFORCE**, an LLM-agent-augmented framework for searching optimal multi-LLM collaboration graphs. Building on REINFORCE, it uses an LLM-based agent to incorporate empirical insights for candidate initialization and distribution updates, following a *sample–feedback–update* pipeline in which feedback serves as textual gradients. The framework comprises three components: the `Agent`, `Archive`, and `Environment`. The `Agent` initializes promising model families and sizes guided by Insight 1 and fixes the best combination within the distribution. In subsequent stages, the new trials are sampled, the `Environment` evaluates them and returns feedback (serving as textual gradients), the `Archive` records the results, and the `Agent` updates the distribution guided by Insights 2 and 3 until convergence. By leveraging LLM-based optimization, our method efficiently identifies graphs that optimize performance alone and graphs that balance with inference latency under joint objectives.

Our **main contributions** are: (i) We study the novel problem of *the search for the optimal multi-LLM collaboration graph for TTS*. (ii) From three empirical insights in multi-LLM collaboration, we develop Agent-REINFORCE, an efficient LLM-guided framework for budget-constrained graph search. (iii) Agent-REINFORCE surpasses traditional and LLM-based baselines in search efficiency and accuracy, and effectively identifies optimal graphs under joint accuracy–latency objectives.

2 Related Work

Test-time Scaling and Compute-optimal Strategy. Allocating additional compute during inference, known as *Test-Time Scaling (TTS)*, can significantly improve LLM performance [56, 55, 5, 58, 44, 75]. TTS methods fall into two main paradigms: *sequential scaling*, which refines outputs iteratively but risks error accumulation, and *parallel scaling*, which aggregates multiple candidates but lacks depth. Hybrid approaches [49, 58] combine both but typically rely on fixed trees and a single model, limiting adaptability. *Compute-optimal TTS* seeks to allocate inference compute most effectively, revealing that small models with optimal strategies might outperform larger ones [5, 58, 65, 49, 53, 7, 21]. Moreover, ensembles of heterogeneous models improve diversity and output quality [27, 2], yet remain underexplored in TTS. Motivated by this gap, we address a novel problem: unifying TTS under a graph structure that enables adaptive topologies and model combinations, and searching for compute-optimal collaboration graphs.

LLMs for Optimization. LLMs, with their rich prior knowledge of machine learning and strong planning ability, have opened new opportunities for practical optimization [74, 19]. Existing research mainly falls into two categories: black-box optimization and hybrid approaches with gradient-based methods. In the black-box setting, LLMs generate and refine candidates using feedback from small training sets [60, 33]. Representative methods include OPRO [60], AgentHPO [33], and GENIUS [76], which leverage task descriptions and prior solution performance for iterative search. LLMs are particularly valuable for initialization, producing high-quality, knowledge-informed solutions that narrow the search space [25, 41]. However, when gradient information is available, black-box approaches become inefficient due to costly evaluations. LLM-based methods address this by interleaving gradient-based training with LLM-guided exploration [19] or by generating textual guidelines as backpropagation signals [66]. Building on these advances, we extend such approaches to compute-optimal test-time scaling by optimizing a probabilistic graph with LLMs for initialization and textual parameter updates. More details of related work are given in Appendix A.18.

3 Preliminaries and Problem Formulation

Test-time Scaling Paradigms and Their Primitives Test-time scaling can be broadly categorized into *parallel scaling* and *sequential scaling*. Given a query q and a language model M with parameters θ, parallel scaling samples k outputs and aggregates them via a fusion function:

$$o = f_{\text{fuse}}(\mathcal{S}, M),\ \mathcal{S} = \{s_i \mid i \in [1, k]\},\ s_i \sim M(s \mid q, \theta). \tag{1}$$

Sequential scaling performs k rounds of self-refinement:

$$o = o^k, \quad o^i = f^i_{\text{refine}}(o^{i-1}, M), \quad o^0 = q. \tag{2}$$

where $f_{\text{fuse}}(\cdot)$ and $f^i_{\text{refine}}(\cdot)$ are both executed by the LLM M, using fusion and refinement prompts, respectively. As shown in Fig. 3, both paradigms can be decomposed into three primitives: repeated sampling, fusion, and self-refinement. Parallel scaling is repeated sampling followed by fusion; sequential scaling is iterative self-refinement. Hybrids recombine these primitives–for example, *Tree-of-Thoughts* [63] uses multi-layer repeated sampling, and *Graph-of-Thoughts* [4] integrates all three primitives in a graph.

Multi-LLM Collaboration Graph for TTS Given the task-specific preference for flexible TTS paradigms beyond the predefined ones, we generalize them into a *multi-LLM collaboration graph* $G = (\mathcal{V}, \mathcal{E}, \mathbf{R}, \mathbf{M})$, where each node $v_i \in \mathcal{V}, i \in [1, n]$, represents an LLM primitive with an assigned role and model, with an example in Fig. 4. Role assignments are denoted by $\mathbf{R} = [r_1, r_2, \ldots, r_n], r_i \in \mathcal{R}$, and model assignments are denoted by $\mathbf{M} = [M_1, M_2, \ldots, M_n], M_i \in \mathcal{M}$. Thus, each node is characterized by a role r_i, which specifies how it processes inputs from its predecessors, and a model M_i, which means which LLM is invoked. Directed edges $e_{ij} \in \mathcal{E}$ represent the flow of information from node v_i to node v_j. We consider two roles $\mathcal{R} = \{\text{assistant}, \text{fuser}\}$, as illustrated in Fig. 4: (i) **Assistant**, which refines the outputs of its predecessors (orange nodes); and (ii) **Fuser**, which aggregates multiple predecessor outputs (green nodes). The collaboration graph G is a directed acyclic graph (DAG) with a designated input node (yellow) that initiates information propagation. Message passing proceeds forward along edges until it reaches a sink node (a node without outgoing edges), whose output serves as the final prediction of the graph.

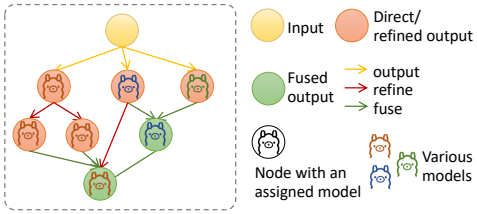

Figure 4: Generalize TTS as a graph.

Inference on Multi-LLM Collaboration Graph As shown in Algo. 2 in Appendix A.2, inference over a multi-LLM collaboration graph G proceeds in topological order. The process begins by identifying the successor nodes of the input node. These nodes process the query to generate initial outputs that are propagated to their successors, reducing the in-degree of their successors by one accordingly. The newly activated nodes (with zero in-degree) are then executed based on their assigned roles and models. A *fuser* aggregates the outputs of its predecessors, whereas an *assistant* refines them. This procedure continues iteratively until all nodes in G have been executed. The output of a unique sink node–node with no outgoing edges, is the final output.

Budget Definition To enable comparative computation across models and topologies, we define the budget using a concrete compute metric, e.g., FLOPs or dollar cost. Let the computational cost of a collaboration graph G be $f_{\text{cost}}(G, T)$. The budget is defined as $B = {}^{f_{\text{cost}}(G,T)}/_{f_{\text{cost}}(G_{\text{smallest}},T)}$, where G_{smallest} is the single-node graph (excluding the input node) using the smallest model, corresponding to one budget unit. Thus, a multi-LLM graph with budget B is equivalent to running B single-node inferences on the smallest model. A detailed introduction to the budget definition is in Appendix A.8. Formally, we report computational cost in FLOPs, which we adopt as our primary cost metric throughout this work.

Proposition 3.1 (FLOPs Cost Function)**.** *For each node v_i, the cost depends on the size of the model and its effective input/output lengths, leading to a dependence on the node in-degree $d(v_i)$. Adding up to all nodes, the total cost can be expressed as* $f_{cost}(G, T) = \sum_{v_i \in \mathcal{V}} \left[\alpha_i \, d(v_i)^2 + \beta_i \, d(v_i) + \gamma_i \right]$,

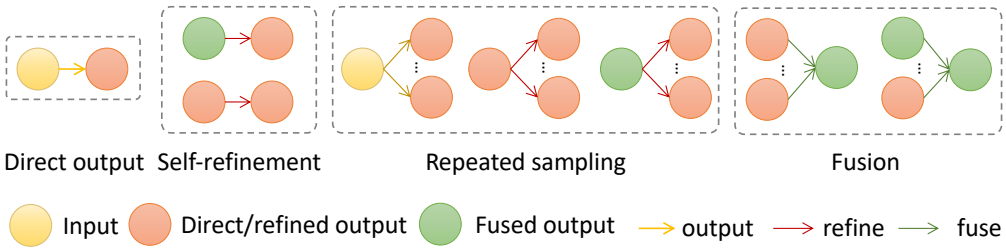

Figure 3: Test-time scaling primitives.

where the coefficients $\alpha_i, \beta_i, \gamma_i$ capture the contributions of the model dimension, depth, and average task input/output lengths. Detailed derivations of $\alpha_i, \beta_i, \gamma_i$ are provided in Appendix A.7.

Problem Definition The goal of *test-time compute-optimal scaling* is to allocate inference compute most effectively under a fixed budget. We formalize this as *searching for the task-specific compute-optimal multi-LLM collaboration graph*. Given training data $\mathcal{D}_{\text{train}}$, test data $\mathcal{D}_{\text{test}}$, a model pool $\mathcal{M} = \{M_1, \ldots, M_n\}$, and a budget B, the objective is to identify a collaboration graph that specifies role and model assignments for nodes, together with the cooperation topology, so as to maximize task performance under the budget constraint. Therefore, our research problem is:

Definition 3.2 (Test-time Compute-optimal Multi-LLM Collaboration Graph for a Task)**.** Given the training set $\mathcal{D}_{\text{train}}$ for a given task T, the model pool $\mathcal{M}$, and a fixed computational budget B, the goal is to find the best collaboration graph that optimizes the performance on $\mathcal{D}_{\text{train}}$, i.e.,

$$G^\star = \arg \max_{G \in \mathcal{G}(\mathcal{M}, B)} u_T(G; \mathcal{D}_{\text{train}}) \tag{3}$$

where $\mathcal{G}(\mathcal{M}, B) = \{G \mid f_{\text{budget}}(G, T) \leq B\}$ is the set of feasible multi-LLM collaboration graphs from $\mathcal{M}$ under budget B. Each $G = (\mathcal{V}, \mathcal{E}, \mathbf{R}, \mathbf{M})$ is a DAG, with node v_i assigned role $r_i \in \{\text{assistant}, \text{fuser}\}$ and model $M_i \in \mathcal{M}$, and edge e_{ij} denoting information flow. The utility function $u_T(G; \mathcal{D}_{\text{train}})$ measures the performance of G on $\mathcal{D}_{\text{train}}$, while $G^\star$ is finally evaluated on $\mathcal{D}_{\text{test}}$.

4 Insights of Multi-LLM Collaboration Graph for TTS

Searching for the optimal multi-LLM collaboration graph for test-time scaling faces two challenges: (i) the search space grows combinatorially with the increased budget, making exhaustive enumeration infeasible; and (ii) the task-specific requirements are highly specific, demanding tailored designs. We therefore conduct pilot experiments to uncover TTS patterns, which pave the way to design a search method for compute-optimal collaboration graphs.

Experimental Setting. We conduct pilot experiments on three tasks: **MATH** [23] (arithmetic reasoning), **MMLU** [22] (general reasoning), and **HumanEval** [8] (code generation), evaluated by accuracy (MATH, MMLU) and pass@1 (HumanEval). The model pool includes LLaMA-3 [1B, 3B, 8B] [16] and Gemma [1B, 2B, 7B] [51]. Details are in Appendix A.5.

Empirical Insights on Model Selection, Parallel and Sequential Scaling. We examine TTS behavior under increasing compute budgets and different model selections, and guide the search for the optimal multi-LLM collaboration graph in Sec.5. Fig. 5 and 6 illustrate how model selection, parallel and sequential scaling, and graph width-depth configuration influence TTS. We get three insights below.

Insight 1: Task-specific preferences for model family and size combinations. We conduct preliminary tests on MATH and MMLU to examine task-specific model preferences. Results in Fig. 5(a–b) show that replicating the strongest model family is generally more effective than mixing families: for example, LLaMA consistently outperforms Gemma in the 3B space on MMLU, so using LLaMA$\times$2 yields higher accuracy than LLaMA+Gemma or Gemma$\times$2. Results in Fig. 5(c–d) show that within a fixed budget, reasoning tasks (MATH) benefit from ensembles of smaller models, while knowledge tasks (MMLU) prefer larger ones. These trends reflect differences in task demands and difficulty: reasoning tasks leverage multiple smaller models for iterative refinement, whereas knowledge tasks require the broader coverage of large models. A more detailed discussion is provided in Appendix A.6. Consequently, **tasks favor replication of the strongest model family, with small-model ensembles preferred only when their incremental gains are substantial.**

Insight 2: Parallel and sequential scaling saturate and decline beyond an optimal budget. Fig. 6(a–b) shows that both parallel (width) and sequential (depth) scaling follow a non-monotonic trend: performance improves up to a task-dependent optimum, then plateaus or declines. On MATH, for example, peak accuracy occurs at 8 parallel or 8 sequential nodes. Beyond these points, added width yields diminishing gains due to long-context limits, while added depth amplifies propagated errors. A more detailed discussion is provided in Appendix A.6. In summary, **both width and depth exhibit task-dependent optima, beyond which extra computation provides negative returns.**

Insight 3: Interdependence between graph width and depth. Fig. 6(c) shows MATH performance under varying width (w) and depth (d) with $wd \leq 24$ using LLaMA-1B. Accuracy rises then falls

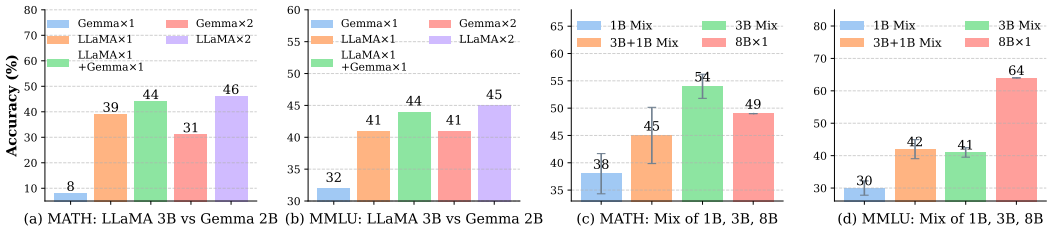

Figure 5: Performance on MATH and MMLU across model family and size. LLaMA by default.

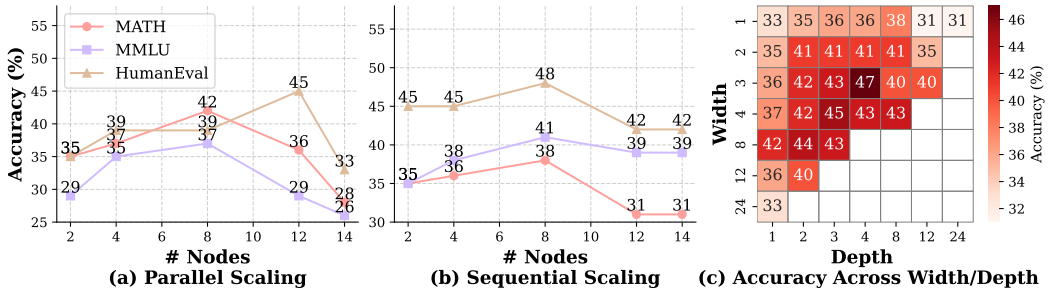

Figure 6: (a–b) Performance with Parallel and Sequential Scaling on various datasets. (c) Heatmap of performance under various Width-Depth collaboration graphs on MATH. Model is LLaMA-3 1B.

as either dimension grows, confirming non-monotonic trends. Moreover, width and depth interact: larger widths reduce the optimal depth (e.g., 8 at w=1 vs. 4 at w=3), while deeper refinement shifts the optimal width forward. A more detailed discussion is in Appendix A.6. Thus, **graph width and depth are interdependent, with growth in one dimension altering the optimum of the other.**

5 Proposed Framework–Agent-REINFORCE

Guided by the insights in Sec. 4, we introduce **Agent-REINFORCE**, an LLM-Agent-augmented REINFORCE algorithm that follows a *sample–feedback–update* loop to find the compute-optimal multi-LLM collaboration graph under a fixed budget. The LLM agent samples candidates and updates graphs using textual feedback (serving the textual gradient in REINFORCE) while integrating task-specific model preferences, budget allocation strategies, and width–depth interactions. We next formalize the probabilistic graph optimization problem and describe our method.

5.1 Probabilistic Graph Optimization Problem

Optimization Problem One way to find the optimal collaboration graph is black-box search, either through enumeration [3] (e.g., grid or random search) or Bayesian optimization [48], which fits a surrogate model to the objective and selects queries via an acquisition function. Yet enumeration is infeasible as the graph space grows exponentially, while standard BO is designed for low-dimensional continuous domains and becomes sample-inefficient in large, discrete spaces. We therefore reformulate the task as a graph optimization problem, leveraging policy-gradient methods for efficient exploration, guided sampling, and budget-aware control. Given a task T and its utility function u_T, let $G \sim \mathbb{P}_{\theta,\pi,\psi}$ denote a sampled multi-LLM collaboration graph. The distribution $\mathbb{P}_{\theta,\pi,\psi}$ is parameterized by three components: $\theta = \{\theta_{ij}\}$, where $\sigma(\theta_{ij}) \in [0,1]$ represents the probability that edge e_{ij} is present; $\pi = \{\pi_i\}$, where $\mathrm{softmax}(\pi_i) \in [0,1]^{|\mathcal{R}|}$ denotes the probability of node v_i selecting a role $r \in \mathcal{R}$; and $\psi = \{\psi_i\}$, where $\mathrm{softmax}(\psi_i) \in [0,1]^{|\mathcal{M}|}$ denotes the probability of node v_i choosing a model $M \in \mathcal{M}$. The optimization problem is to identify

$$\theta^\star, \pi^\star, \psi^\star = \arg\max_{\theta,\pi,\psi} \mathbb{E}_{G\sim\mathbb{P}_{\theta,\pi,\psi}}\left[u_T(G, D_{\text{train}})\right] \quad \text{s.t.} \quad f_{\text{budget}}(G,T) \le B. \tag{4}$$

Algorithm 1 Agent-REINFORCE: Compute-Optimal Collaboration Graph Optimization

Require: task T, model set $\mathcal{M}$, agent $\mathcal{A}$, environment $\mathcal{E}$, budget B
Ensure: optimized collaboration graph G

Initialize archive $\mathcal{L} \leftarrow \emptyset$
Stage 1: $\mathcal{C} \leftarrow \mathcal{A}.\texttt{select_family_size}(T, \mathcal{M}, B)$; $\mathcal{S} \leftarrow \mathcal{E}.\texttt{execute}(\mathcal{C})$ (Init. Stage 1)
Stage 2: $\mathcal{C} \leftarrow \mathcal{A}.\texttt{select_instance}(T, \mathcal{M}, \mathcal{S}, B)$; $\mathcal{S} \leftarrow \mathcal{E}.\texttt{execute}(\mathcal{C})$ (Init. Stage 2)
Initialize nodes in $\tilde{\mathbf{G}}$ with the best model family, size, and instance count (Insight 1)
repeat
  Update archive $\mathcal{L} \leftarrow \mathcal{L} \cup \{(\mathcal{C}, \mathcal{S}, \tilde{\mathbf{G}})\}$
  Sample new trials $\mathcal{C} \leftarrow \mathcal{A}.\texttt{sampling}(\tilde{\mathbf{G}}, B)$
  Get feedback (textual gradient) $\mathcal{S} \leftarrow \mathcal{E}.\texttt{execute}(\mathcal{C})$
  Update graph $\tilde{\mathbf{G}} \leftarrow \mathcal{A}.\texttt{update}(\mathcal{C}, \mathcal{S}, \mathcal{L}, \tilde{\mathbf{G}})$ (Insights 2,3)
until stopping criterion is met
return graph G by deterministic decoding from $\tilde{\mathbf{G}}$

5.2 Agent-REINFORCE

The REINFORCE algorithm [57] can optimize Eq.(4) via gradient ascent through iterative *sample–gradient–update* (sampling candidates, estimating gradients from their utility, and updating parameters; see Appendix A.11 for details). However, its step-by-step updates lead to slow progress, local optima, and difficulty in incorporating prior insights or semantic knowledge. To overcome these limitations, we propose **Agent-REINFORCE**, an LLM-agent-augmented framework which builds on REINFORCE but replaces gradients with feedback-conditioned updates. Each iteration follows a *sample–feedback–update* loop: guided by empirical insights, the agent samples candidate graphs, receives feedback as textual gradients, and updates the distribution iteratively until convergence. As shown in Fig. 7(b–d), the framework comprises three components: `Agent`, `Archive`, and `Environment`. The `Agent` first generates candidate trials of the model family and size combinations (guided by Insight 1). Feedback from the `Environment` selects the best model assignments and initializes the probabilistic graph distribution. In subsequent iterations, the `Agent` samples new trials from the updated distribution $\mathbb{P}_{\theta,\pi,\psi}$, the `Environment` tests them, and the `Archive` records results. The `Agent` then updates the distribution based on feedback and history, and this loop continues until convergence. The procedure is given in Algo. 1.

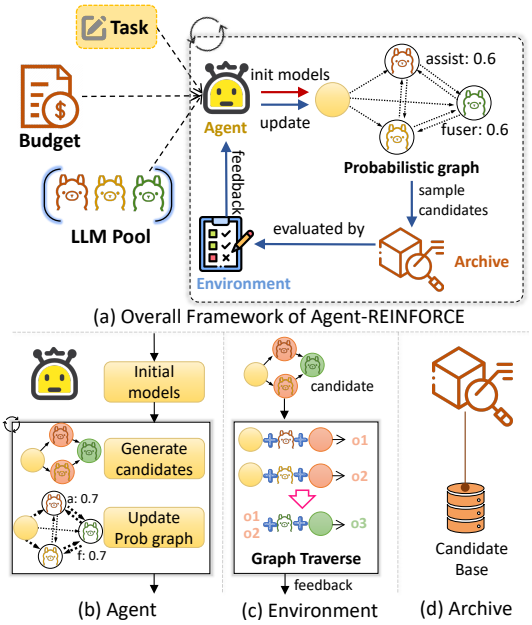

Figure 7: Overview of Agent-REINFORCE for Optimizing Collaboration Graph.

Agent component. The LLM-based `Agent`, in Fig. 7 (b), initializes model assignments, samples new trials, and updates the probabilistic graph. Since LLMs lack prior knowledge of test-time scaling, which is relatively new, we incorporate Insight 1 to guide the initialization of model assignments, and Insights 2 and 3 to inform subsequent updates. *Insight 1* shows that tasks prefer replicating the strongest family, with small-model ensembles chosen when their gains are high. Hence, initialization focuses on task-specific model assignments (family, size, and instances) to guide optimization and reduce wasted exploration. We initialize the family-size and instance counts in two stages.

First, the `Agent` identifies family and size preferences using each model's meta-information from HuggingFace [1], including prior performance and the task description. Prior performance guides family selection; when unavailable, initial trials pre-test each model's prior performance to infer family preferences. For size selection, the incremental gains from ensembling one versus two small models relative to a single large model inform size preference, motivating trials that explore both small ensembles

and large models. Therefore, the agent initializes candidates as $\mathcal{A}$.`select_family_size`$(T, \mathcal{M}, B)$, retaining only those within budget B, and obtains performance scores from the `Environment` as feedback $\mathcal{S}$ to identify the preferred family and size (Algo. 1, Line 2).

Second, using feedback $\mathcal{S}$, the `Agent` generates diverse candidate model combinations within budget B via $\mathcal{A}$.`select_instance`$(T, \mathcal{M}, \mathcal{S}, B)$, prioritizing the selected family and size while varying instance counts. For each candidate, graph topologies and role assignments are randomly sampled (Algo. 1, Line 3). Feedback is averaged, and the best configuration, covering family, size, and instances, initializes the graph (Algo. 1, Line 4). In subsequent stages, nodes retain the model assignments, while edges and roles are sampled from the probabilistic graph $\tilde{\mathbf{G}}$ via $\mathcal{A}$.`sampling`$(\tilde{\mathbf{G}}, B)$.

Insight 2 shows that width and depth have task-specific optima: performance improves with more nodes up to a point, then degrades. We incorporate this into the update prompt ($\mathcal{A}$.`update`, Algo. Line 9), guiding the `Agent` to "identify the optimization direction for finding the optimal width and depth" by leveraging feedback from current and past trials to adjust the probabilistic graph toward the optimal width–depth balance and accelerate convergence.

Insight 3 highlights the interdependence between width and depth: under a fixed budget, improving one often requires reducing the other. To manage this, we embed an instruction into the update prompt ($\mathcal{A}$.`update`, Algo. Line 9) that directs the `Agent` to exploit the LLM's planning ability to explore these trade-offs between width and depth and adaptively identify critical graphs within budget.

The instructions derived from the insights are applied continuously during the optimization process. Based on the feedback, the `Agent` updates the probabilistic graph (Algo. Line 9), which is then used to sample the next batch of trials (Algo. Line 7). Since each sampled compute graph's structure affects the used budget (denser graphs require more budget), some may exceed the budget limit B. In such cases, we remove the smallest nodes to meet the budget. Correspondingly, the probabilistic graph is updated by removing the same nodes, ensuring that all sampled candidates satisfy the budget constraint. The prompt design is provided in Appendix A.12.

Environment & Archive Components. `Environment` converts candidate graphs from the `Agent` into executable scripts, runs them in the actual task platform on a small training batch, and returns performance feedback (Fig. 7c; Algo. 1, Lines 2–3,8). `Archive` stores the probabilistic graph, trials, and feedback (Fig. 7d; Algo. 1, Lines 1,6), tracking the optimization process across iterations and providing historical traces for the `Agent` to refine updates.

6 Experiments

This section analyzes Agent-REINFORCE for compute-optimal collaboration graphs in TTS, covering ablations, varying budgets, joint objectives, alternative budget metrics, and visualizations.

Experimental Setup. We experiment on MATH, MMLU, and HumanEval using LLaMA models (1B-8B) [16] and Gemma models (1B-7B) [51] (details in Appendix A.5). Baselines fall into three groups: (i) traditional: Bayesian Optimization (BO) [28, 48] and random search; (ii) gradient-based: GPTSwarm [77], a REINFORCE framework with gradient updates, and MaaO [19], combining gradient training with LLM guidance; and (iii) LLM-based: TextGrad [66], which relies solely on textual guidelines. As these methods are not tailored to our setting, we adapt them for test-time compute-optimal graph search (details in Appendix A.13). All methods are run for up to 30 search iterations on the training data and use the validation set to determine convergence. Search is stopped if the average validation performance does not improve for 10 iterations. The final searched graph is evaluated on the test set. We use DeepSeek-R1 [18] as the LLM search agent. We use 18, 42, and 80 as low, medium, and high budgets, roughly matching $1\times$, $3\times$, and $5\times$ largest-model inferences.

Main Results. Tab. 1 reports test performance and convergence time, and Fig. 9 in the Appendix shows training trajectories. We observe: (1) Our method achieves the highest average test-set score (higher accuracy or Pass@1) while converging substantially faster (lower search time). This is enabled by Insights 2–3, which guide the search toward promising regions, and Insight 1, which provides a strong initialization and avoids wasted trials. (2) Compared with the LLM-based TextGrad, our method is much more efficient by pruning high-latency candidates early. Among the methods, TextGrad yields the highest inference latency in the searched graphs, reflecting its tendency to favor dense connections or larger node counts that drive full-budget utilization. Such usage often produces high-overhead graphs and consequently slower convergence. (3) The gradient-based GPTSwarm

Table 1: Performance across MATH, MMLU, and HumanEval at 80 budget. Acc (↑) means Accuracy (%, higher is better), Sear. (↓) means total search time in seconds (lower is better), and Inf. (↓) means average inference time in seconds per test query (lower is better). Best in each column is bolded.

Method	MATH			MMLU			HumanEval			Average		
	Acc	Sear.	Inf.	Acc	Sear.	Inf.	P@1	Sear.	Inf.	Sco.	Sear.	Inf.
Random	39	2852	27.3	44	658	17.8	63	1560	12.6	49	1690	19.2
BO [48]	42	3076	38.0	38	2150	26.8	33	2588	42.6	38	2605	35.8
GPTSwarm [77]	40	943	18.4	42	**463**	10.6	55	804	15.5	46	737	14.8
MaaO [19]	34	1440	24.6	41	738	8.1	42	860	20.7	39	1013	17.8
TextGrad [66]	41	3687	40.4	46	2276	9.6	42	2842	17.3	43	2935	22.4
Ours	**56**	**804**	**17.3**	**54**	493	**8.0**	**73**	**300**	**4.9**	**61**	**532**	**10.0**

and MaaO converge quickly but often produce graphs inferior even to random search, due to their vulnerability to local optima. This underscores the importance of combining global exploration with local refinement. (4) The traditional Bayesian optimization method also suffers from local optima and slow updates due to a lack of task-specific guidance. Random search shows some robustness and can occasionally find competitive solutions, but it remains inefficient and unstable. We report the detailed LLM-agent overhead in Appendix A.15, showing the cost is mainly from iterative trial generation.

Ablation Studies. We evaluate the contribution of each insight through ablation, comparing the full method with variants: `w/o Insight 1` uses random initialization instead of task- and model-informed initialization, while `w/o Insight 2/3` removes prompt components for budget optima and width–depth dependencies. Tab. 2 shows that removing insights generally slows convergence by generating inefficient graphs; `w/o Insight 1` enlarges the candidate space, and `w/o Insight 2/3` biases exploration toward high-budget graphs. Performance drops most under `w/o Insight 1`, as random initialization yields suboptimal starts that limit later search. Excluding Insight 2 or 3 also reduces accuracy by losing guidance on budget and width–depth trade-offs. We also perform an ablation by removing role setting, letting all nodes process predecessors' outputs and generate new answers, which degrades graph performance on MATH. This highlights the importance of the *fuser–assistant* role division in test-time scaling. We note that MMLU performance remains stable without Insight 3 or role settings, as it favors larger models with fewer nodes, reducing the impact of width–depth trade-offs and roles.

Table 2: Ablation study of Agent-REINFORCE on MATH and MMLU w/o insights and role setting.

Methods	MATH		MMLU	
	Acc	Sear.	Acc	Sear.
Agent-REINFORCE	56	804	54	493
w/o Insight 1	45	1946	42	1293
w/o Insight 2	49	2208	47	896
w/o Insight 3	48	1436	54	487
w/o Role	52	785	54	677

Generality of the insights. To test whether the proposed insights generalize beyond our method, we further incorporate them into two baselines, GPTSwarm and MaaO, under the same budget setting as in the experiments section. Specifically, we apply our insight-guided model initialization strategy and introduce an edge regularizer into their graph optimization process. In Tab. 3, these insights consistently improve both the baselines on MATH and MMLU, suggesting that the gains are not specific to our optimizer. In addition, we compare against the structured evolutionary search baseline AFlow [69] to highlight the value of our method. Nevertheless, our full method still achieves the best performance. This is likely because numerical gradient optimization is more prone to local optima in the discrete collaboration-graph space, while structured evolutionary search baselines such as AFlow are designed for general workflow construction and do not explicitly account for the characteristics of test-time scaling.

Table 3: Effect of incorporating our insights into GPTSwarm and MaaO.

Method	MATH	MMLU
GPTSwarm	40	42
GPTSwarm + insights	45	53
MaaO	34	41
MaaO + insights	53	51
AFlow	45	40
Ours	56	54

Performance Under Various Budget Settings. We evaluate search performance on the MATH dataset under FLOPs budgets of 18 and 42, accommodating 1×8B and $[2,3]\times8$B models, respectively. As shown in Tab. 4, our method consistently delivers superior efficiency and accuracy, demonstrating strong generalization. Notably, some baselines perform better at smaller budgets (e.g., MaaO: 44 at budget 18 vs. 34 at 80) because they overlook that the optimal budget is often below the maximum.

Table 4: MATH Acc, Sear, and Inf under various FLOPs and price budgets.

Method	Price $\leq$ \$5E-4			FLOPs Budget 42			FLOPs Budget 18		
	Acc	**Sear.**	**Inf.**	**Acc**	**Sear.**	**Inf.**	**Acc**	**Sear.**	**Inf.**
Random	35	2546	52.5	33	1706	56.9	39	1440	16.0
BO	36	2372	56.6	45	2724	49.5	38	1634	23.1
GPTSwarm	43	832	20.5	44	858	31.2	44	1028	29.1
MaaO	47	1104	20.0	46	889	50.9	44	836	14.6
TextGrad	22	3062	57.9	45	2661	48.6	40	2553	16.8
Ours	**56**	648	18.1	**50**	726	23.8	**47**	771	16.7
Ours (joint)	52	**415**	**14.0**	47	**698**	**17.4**	44	**717**	**5.6**

As noted in Insight 2, computation beyond the optimum yields negative returns, whereas smaller budgets closer to the budget optimum can bring these methods nearer to peak performance.

Latency-aware Joint Optimization Objective. To demonstrate our method's ability to handle joint optimization objectives, we optimize both performance and latency through multidimensional feedback, achieving a balance between accuracy and efficiency. The details of the optimization with a joint objective are in Appendix A.9. As shown in Tab.4, on MATH with an 18 FLOPs budget, the searched graph achieves an average latency of 5.6 seconds per test query, which is much lower than the 16.7 seconds under a performance-only objective, thereby validating its effectiveness for multi-objective optimization, even though performance decreases slightly from 47 to 44.

Generalization to the Dollar Cost as Budget. Beyond FLOPs, end-users often care about the monetary cost of API calls. We introduce price as an additional budget metric, directly measured in currency units. As shown in Tab. 6 (Appendix A.10) , cost scales with input and output tokens, so $f_{\text{budget}}(G, T)$ is redefined as input length times per-token input price plus output length times per-token output price. Under a fixed API budget \$5E-4 per query (from $4\times$8B to $6\times$8B models), the results in Tab. 4 show Agent-REINFORCE excels in both accuracy and efficiency, showing strong generalization across cost metrics.

Visualization Fig. 8 shows the optimal collaboration graph under a budget of 80 for the MATH task. The graph favors small-model ensembles, as the relatively low task difficulty allows small models to achieve strong performance, with additional instances further improving effectiveness. It favors a hybrid scaling pattern biased toward sequential refinement (width 3, depth 4), reflecting the advantage of iterative self-refinement in multi-step mathematical reasoning.

7 Conclusion

We study a novel problem of task-specific, compute-optimal test-time scaling over multi-LLM collaboration graphs under a fixed budget, where the design space grows exponentially. Pilot analysis reveals three insights: (1) tasks replicate the strongest model family, favoring small-model ensembles when marginal gains are high; (2) width and depth admit task-specific optima, beyond which additional compute degrades performance; and (3) width and depth are coupled, with growth in one shifting the optimum of the other. Based on these findings, we propose **Agent-REINFORCE**, a budget-aware, feedback-driven LLM-agent framework for collaboration graph search. Experiments show that it outperforms traditional and LLM-based baselines in both search efficiency and final performance, and identifies optimal graphs under a joint performance–latency objective.

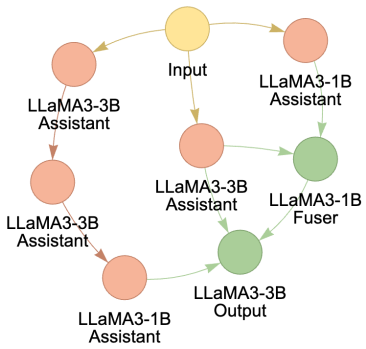

Figure 8: Optimal graph on MATH.

Limitations: While our TTS method achieves promising results, we acknowledge several areas for potential future exploration. We provide a detailed discussion in Appendix A.19.

References

[1] Hugging face. https://huggingface.co/. Accessed: 2025-09-22.

[2] Mari Ashiga, Wei Jie, Fan Wu, Vardan Voskanyan, Fateme Dinmohammadi, Paul Brookes, Jingzhi Gong, and Zheng Wang. Ensemble learning for large language models in text and code generation: A survey. *arXiv preprint arXiv:2503.13505*, 2025.

[3] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The journal of machine learning research*, 13(1):281–305, 2012.

[4] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690, 2024.

[5] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

[6] Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Alphamath almost zero: Process supervision without process. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=VaXnxQ3UKo.

[7] Hao Mark Chen, Zhiwen Mo, Guanxi Lu, Shuang Liang, Lingxiao Ma, Wayne Luk, and Hongxiang Fan. Fasttts: Accelerating test-time scaling for edge llm reasoning. In *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 732–748, 2026.

[8] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[9] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *ICLR*, 2024.

[10] Weizhe Chen, Sven Koenig, and Bistra Dilkina. Iterative deepening sampling for large language models. *arXiv e-prints*, pages arXiv–2502, 2025.

[11] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KuPixIqPiq.

[12] Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. LM vs LM: Detecting factual errors via cross examination. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12621–12640, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.778. URL https://aclanthology.org/2023.emnlp-main.778/.

[13] I De Zarzà, J De Curtò, Gemma Roig, and Carlos T Calafate. Optimized financial planning: integrating individual and cooperative budgeting models with llm recommendations. *AI*, 5(1): 91–114, 2023.

[14] Shangbin Feng, Zifeng Wang, Palash Goyal, Yike Wang, Weijia Shi, Huang Xia, Hamid Palangi, Luke Zettlemoyer, Yulia Tsvetkov, Chen-Yu Lee, et al. Heterogeneous swarms: Jointly optimizing model roles and weights for multi-llm systems. *arXiv preprint arXiv:2502.04510*, 2025.

[15] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Sx038qxjek.

[16] 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.

[17] Lin Gui, Cristina Garbacea, and Victor Veitch. BoNBon alignment for large language models and the sweetness of best-of-n sampling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=haSKMlrbX5.

[18] 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.

[19] Zixian Guo, Ming Liu, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Llm as a complementary optimizer to gradient descent: A case study in prompt tuning. *arXiv preprint arXiv:2405.19732*, 2024.

[20] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, 2023.

[21] Zixu Hao, Jianyu Wei, Tuowei Wang, Minxing Huang, Huiqiang Jiang, Shiqi Jiang, Ting Cao, and Ju Ren. Scaling llm test-time compute with mobile npu on smartphones. In *Proceedings of the 21st European Conference on Computer Systems*, pages 2157–2172, 2026.

[22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[23] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

[24] Xuhan Huang, Qingning Shen, Yan Hu, Anningzhe Gao, and Benyou Wang. LLMs for mathematical modeling: Towards bridging the gap between natural and mathematical languages. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2678–2710, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.146. URL https://aclanthology.org/2025.findings-naacl.146/.

[25] Ganesh Jawahar, Muhammad Abdul-Mageed, Laks Lakshmanan, and Dujian Ding. Llm performance predictors are good initializers for architecture search. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10540–10560, 2024.

[26] Zipeng Ji, Guanghui Zhu, Chunfeng Yuan, and Yihua Huang. RZ-NAS: Enhancing LLM-guided neural architecture search via reflective zero-cost strategy. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9UExQpH078.

[27] 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.

[28] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13:455–492, 1998.

[29] Shiyuan Li, Yixin Liu, Qingsong Wen, Chengqi Zhang, and Shirui Pan. Assemble your crew: Automatic multi-agent communication topology design via autoregressive graph generation. *arXiv preprint arXiv:2507.18224*, 2025.

[30] Zichong Li, Xinyu Feng, Yuheng Cai, Zixuan Zhang, Tianyi Liu, Chen Liang, Weizhu Chen, Haoyu Wang, and Tuo Zhao. Llms can generate a better answer by aggregating their own responses. *arXiv preprint arXiv:2503.04104*, 2025.

[31] Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025.

[32] Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling. *arXiv preprint arXiv:2502.06703*, 2025.

[33] Siyi Liu, Chen Gao, and Yong Li. Large language model agent for hyper-parameter optimization. *arXiv preprint arXiv:2402.01881*, 2024.

[34] Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance bayesian optimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OOxotBmGol.

[35] Yexiang Liu, Zekun Li, Zhi Fang, Nan Xu, Ran He, and Tieniu Tan. Rethinking the role of prompting strategies in LLM test-time scaling: A perspective of probability theory. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27962–27994, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1356. URL https://aclanthology.org/2025.acl-long.1356/.

[36] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic LLM-agent network: An LLM-agent collaboration framework with agent team optimization, 2024. URL https://openreview.net/forum?id=i43XCU54Br.

[37] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

[38] Parsa Mirtaheri, Ezra Edelman, Samy Jelassi, Eran Malach, and Enric Boix-Adsera. Let me think! a long chain-of-thought can be worth exponentially many short ones. *arXiv preprint arXiv:2505.21825*, 2025.

[39] Kou Misaki, Yuichi Inoue, Yuki Imajuku, So Kuroki, Taishi Nakamura, and Takuya Akiba. Wider or deeper? scaling LLM inference-time compute with adaptive branching tree search. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. URL https://openreview.net/forum?id=3HF6yogDEm.

[40] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[41] Yves Gaetan Nana Teukam, Federico Zipoli, Teodoro Laino, Emanuele Criscuolo, Francesca Grisoni, and Matteo Manica. Integrating genetic algorithms and language models for enhanced enzyme design. *Briefings in bioinformatics*, 26(1):bbae675, 2025.

[42] Muhammad Umair Nasir, Sam Earle, Julian Togelius, Steven James, and Christopher Cleghorn. Llmatic: neural architecture search via large language models and quality diversity optimization. In *proceedings of the Genetic and Evolutionary Computation Conference*, pages 1110–1118, 2024.

[43] Jiayi Pan, Xiuyu Li, Long Lian, Charlie Victor Snell, Yifei Zhou, Adam Yala, Trevor Darrell, Kurt Keutzer, and Alane Suhr. Learning adaptive parallel reasoning with language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=YgwQ7sXPXU.

[44] Qianjun Pan, Wenkai Ji, Yuyang Ding, Junsong Li, Shilian Chen, Junyi Wang, Jie Zhou, Qin Chen, Min Zhang, Yulan Wu, et al. A survey of slow thinking-based reasoning llms using reinforcement learning and test-time scaling law. *Information Processing & Management*, 63 (2):104394, 2026.

[45] Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large language model-based multi-agent collaboration. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=K3n5jPkrU6.

[46] Jon Saad-Falcon, Adrian Gamarra Lafuente, Shlok Natarajan, Nahum Maru, Hristo Todorov, Etash Guha, E Kelly Buchanan, Mayee Chen, Neel Guha, Christopher Ré, et al. Archon: An architecture search framework for inference-time techniques. *arXiv preprint arXiv:2409.15254*, 2024.

[47] Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. Scaling test-time compute without verification or RL is suboptimal. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=beeNgQEfe2.

[48] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104 (1):148–175, 2015.

[49] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

[50] Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. Fast best-of-n decoding via speculative rejection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=348hfcprUs.

[51] Gemma Team. Gemma 3. 2025. URL https://arxiv.org/abs/2503.19786.

[52] Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. Alphazero-like tree-search can guide large language model decoding and training. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=C4OpREezgj.

[53] Fali Wang, Hui Liu, Zhenwei Dai, Jingying Zeng, Zhiwei Zhang, Zongyu Wu, Chen Luo, Zhen Li, Xianfeng Tang, Qi He, et al. Agenttts: Large language model agent for test-time compute-optimal scaling strategy in complex tasks. *arXiv preprint arXiv:2508.00890*, 2025.

[54] Tianchun Wang, Zichuan Liu, Yuanzhou Chen, Jonathan Light, Haifeng Chen, Xiang Zhang, and Wei Cheng. Diversified sampling improves scaling llm inference. *arXiv preprint arXiv:2502.11027*, 2025.

[55] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

[56] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[57] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[58] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The Thirteenth International Conference on Learning Representations*, 2025.

[59] Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36:41618–41650, 2023.

[60] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bb4VGOWELI.

[61] Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*, 2025.

[62] Yingxuan Yang, Huacan Chai, Shuai Shao, Yuanyi Song, Siyuan Qi, Renting Rui, and Weinan Zhang. Agentnet: Decentralized evolutionary coordination for llm-based multi-agent systems. *arXiv preprint arXiv:2504.00587*, 2025.

[63] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[64] Fei Yu, Anningzhe Gao, and Benyou Wang. Ovm, outcome-supervised value models for planning in mathematical reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 858–875, 2024.

[65] Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. Inference scaling for long-context retrieval augmented generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FSjIrOm1vz.

[66] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

[67] Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? *arXiv preprint arXiv:2502.12215*, 2025.

[68] Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-designer: Architecting multi-agent communication topologies via graph neural networks. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=LpE54NUnmO.

[69] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=z5uVAKwmjf.

[70] Lei Zhang, Yuge Zhang, Kan Ren, Dongsheng Li, and Yuqing Yang. Mlcopilot: Unleashing the power of large language models in solving machine learning tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2931–2959, 2024.

[71] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. What, how, where, and how well? a survey on test-time scaling in large language models. *CoRR*, 2025.

[72] Shun Zhang, Zhenfang Chen, Yikang Shen, Mingyu Ding, Joshua B. Tenenbaum, and Chuang Gan. Planning with large language models for code generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Lr8cOOtYbfL.

[73] Tuo Zhang, Jinyue Yuan, and Salman Avestimehr. Revisiting OPRO: The limitations of small-scale LLMs as optimizers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1727–1735, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.100. URL https://aclanthology.org/2024.findings-acl.100/.

[74] Yisong Zhang, Ran Cheng, Guoxing Yi, and Kay Chen Tan. A systematic survey on large language models for evolutionary optimization: From modeling to solving. *arXiv preprint arXiv:2509.08269*, 2025.

[75] Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, et al. Genprm: Scaling test-time compute of process reward models via generative reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34932–34940, 2026.

[76] Mingkai Zheng, Xiu Su, Shan You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. Can gpt-4 perform neural architecture search? *arXiv preprint arXiv:2304.10970*, 2023.

[77] Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. Gptswarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*, 2024.

Algorithm 2 Inference on Multi-LLM Collaboration Graph for TTS

Require: query q, graph $G = (\mathcal{V}, \mathcal{E}, \mathbf{R}, \mathbf{M})$ (a DAG with unique sink v_{sink})
Ensure: final output o

Initialize $d_{\text{in}}(v)$, $d_{\text{out}}(v)$ and buffers $\mathcal{O}(v) \leftarrow \emptyset$ for all $v \in \mathcal{V}$
$\mathcal{Q} \leftarrow \{ v \in \mathcal{V} \mid d_{\text{in}}(v) = 0 \}$ (topological frontier)
while $\mathcal{Q} \neq \emptyset$ **do**
  Remove a node v from $\mathcal{Q}$
  $\mathcal{C} \leftarrow \bigcup_{u \in \text{pred}(v)} \mathcal{O}(u)$
  if $r_v =$ fuser **then**
    $\mathcal{O}(v) \leftarrow f_{\text{fuse}}(q, \mathcal{C}, M_v)$
  else
    $\mathcal{O}(v) \leftarrow f_{\text{refine}}(q, \mathcal{C}, M_v)$ ($r_v =$ assistant)
  end if
  for $w \in \text{succ}(v)$ **do**
    $d_{\text{in}}(w) \leftarrow d_{\text{in}}(w) - 1$
    if $d_{\text{in}}(w) = 0$ **then**
      Add w to $\mathcal{Q}$
    end if
  end for
end while
return $o \leftarrow \mathcal{O}(v_{\text{sink}})$ (unique sink with $d_{\text{out}}(v_{\text{sink}}) = 0$)

A Appendix

A.1 Test-Time Scaling: Modes and Building Blocks

Fig. 2 shows four paradigms: (a) parallel via repeated sampling + aggregation; (b) sequential via iterative self-refinement; (c) fixed hybrids that fuse both; and (d) our dynamic setting that searches architectures and assigns heterogeneous models under a compute budget. Fig. 3 reduces these to three primitives—repeated sampling, fusion, self-refinement, and frames dynamic TTS as a multi-LLM collaboration graph with role-assigned nodes (e.g., fuser, assistant), directed information flow, and a terminal aggregator.

A.2 Inference on Multi-LLM Collaboration Graph for TTS Algorithm

Algo 2 executes the collaboration graph G in topological order: Successors of the input node generate initial outputs; nodes activate when in-degree reaches zero and run by role—*fuser* (aggregate) or *assistant* (refine)—propagating results forward. The unique sink node produces the final answer.

A.3 Pilot Experiments for Existing TTS

Table 5 summarizes the task-specific preferences for topologies and model combinations. In MATH, the hybrid graph topology combined with a mixture of 3B models yields the best accuracy. In contrast, MMLU shows a clear preference for pure parallel graph topologies and the use of a single 8B model. These results indicate that different tasks exhibit distinct preferences for architectural patterns and model configurations.

A.4 Calculation of the number of DAGs

Given n nodes, the spectrum of possible configurations ranges from *totally indistinguishable* nodes to *totally distinguishable* nodes. The number of directed acyclic graphs (DAGs) lies within this range: the indistinguishable case corresponds to counting the number of *non-isomorphic DAGs* (where isomorphic topologies are counted only once), while the distinguishable case corresponds to counting the number of *labeled DAGs*.

Indistinguishable nodes: Number of non-isomorphic DAGs. A closed-form characterization can be derived from the fact that every DAG admits at least one topological ordering. If we fix the order

Table 5: Accuracy (ACC, %) across different topologies and model combinations on MATH and MMLU. LLaMA-3 models are used by default. Results are averaged over 10 random graphs.

Dataset	Model Comb.	Sequential	Parallel	Hybrid
MATH	1B Mix	37	41	44
	3B Mix	58	56	59
	8B×1	49	49	49
MMLU	1B Mix	35	43	30
	3B Mix	49	51	41
	8B×1	64	64	64

$1 < 2 < \cdots < n$, then only edges of the form $i \to j$ with $i < j$ are permitted. This yields $\frac{n(n-1)}{2}$ possible edges, and thus $2^{\binom{n}{2}}$ candidate adjacency matrices, all acyclic by construction. However, many of these candidates are *isomorphic*. To correctly count non-isomorphic DAGs, each candidate graph is reduced to a *canonical labeling*, and graphs with the same canonical form are merged. To reduce the cost of considering all permutations, nodes are grouped by their in-degree and out-degree, and permutations are applied only within these groups, which substantially reduces computational complexity. The Python implementation in Listing 1 computes the number of non-isomorphic DAGs.

Distinguishable nodes: Number of labeled DAGs. When nodes are labeled, the total number of DAGs can be computed using a well-known recurrence relation: $A(0) = 1, \quad A(n) = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} 2^{k(n-k)} A(n-k)$. Here, $A(n)$ denotes the number of labeled DAGs on n nodes. This formulation accounts for all possible edge configurations under node labeling and ensures that only acyclic structures are counted. The corresponding Python implementation is provided in the Listing 2.

```python
import itertools as it

def upper_adj_bitmasks(n, bits):
    rows = [0]*n
    for i in range(n):
        for j in range(i+1, n):
            if bits & 1: rows[i] |= (1<<j)
            bits >>= 1
    return rows

def indegree_outdegree(rows):
    n = len(rows)
    outdeg = [r.bit_count() for r in rows]
    indeg = [0]*n
    for i,r in enumerate(rows):
        while r:
            j = (r&-r).bit_length()-1
            indeg[j] += 1
            r &= r-1
    return tuple(zip(outdeg, indeg))

def permute_rows(rows, perm):
    inv = [0]*len(perm)
    for i,p in enumerate(perm): inv[p]=i
    return [sum(1<<inv[j] for j in range(len(rows)) if (rows[perm[i]]>>j)&1) for i in range(
        len(rows))]

def canonical_form_rows(rows):
    degs = indegree_outdegree(rows)
    groups = {}
    for i,deg in enumerate(degs): groups.setdefault(deg,[]).append(i)
    perms = [it.permutations(g) for g in groups.values()]
    best = None
    for p in it.product(*perms):
        perm = [x for part in p for x in part]
        newrows = permute_rows(rows, perm)
        key = ''.join('1' if (newrows[i]>>j)&1 else '0' for i in range(len(rows)) for j in
            range(len(rows)))
        if best is None or key < best: best = key
    return best

def count_unlabeled_dags(n):
```

```
    m = n*(n-1)//2
    return len({canonical_form_rows(upper_adj_bitmasks(n,b)) for b in range(1<<m)})

for n in range(1,8): print(n, count_unlabeled_dags(n))

# Results (number of non-isomorphic DAGs)
# n=1: 1
# n=2: 2
# n=3: 8
# n=4: 54
# n=5: 762
# n=6: 21,542
# n=7: 1,259,209
```

Listing 1: Python code computes the number of non-isomorphic DAGs

```
import math
from functools import lru_cache

@lru_cache(None)
def labeled_dags(n):
    if n==0:
        return 1
    s=0
    for k in range(1,n+1):
        s += (-1)**(k+1) * math.comb(n,k) * (2**(k*(n-k))) * labeled_dags(n-k)
    return s

for n in range(1,9):
    print(n, labeled_dags(n))

# n=8: 783,702,329,343
```

Listing 2: Python code computes the number of labeled DAGs

A.5 Tasks, Datasets, and Models

MATH dataset [23] The MATH dataset is used for *arithmetic reasoning* evaluation tasks, consisting of 12,500 competition-level problems from high school contests. Each problem is accompanied by a step-by-step solution, which supports evaluation of final-answer *accuracy* as the primary metric. Serving as a rigorous benchmark for symbolic manipulation and multi-step mathematical reasoning, MATH is widely used to test the limits of language models. In our experiments, we sample 750 problems for training and 100 for testing, with average prompt and generation lengths of 202 and 275 tokens, respectively.

MMLU dataset [22] The Massive Multitask Language Understanding (MMLU) dataset is a comprehensive benchmark for evaluating *knowledge and general reasoning tasks* across 57 tasks spanning humanities, social sciences, STEM, and professional fields, with questions ranging from elementary to advanced difficulty. Each task is presented in a multiple choice format and *precision* is used as a standard evaluation metric. MMLU has become a widely adopted benchmark for assessing the general knowledge and cross-domain adaptability of large language models. In our experiments, we randomly sampled 285 questions for training and 100 for testing, with average prompt and generation lengths of 213 and 230 tokens, respectively.

HumanEval dataset [8] The HumanEval dataset is a benchmark designed to assess *code generation and synthesis* capabilities of language models. It contains 164 Python programming problems, each consisting of a function signature, natural language docstring, and unit tests for automatic evaluation. The primary metric is *passk*, which measures the probability that at least one of k generated solutions passes all hidden test cases. HumanEval has become a standard benchmark for evaluating the ability of models to translate natural language descriptions into correct, executable code. In our experiments, we randomly sample 128 instances for training and others for testing, with average prompt and generation lengths of 181 and 104 tokens, respectively.

Language models adopted We evaluate our method using language models of varying scales from the LLaMA-3 family [16] and Gemma family [51]. To promote diversity in generations and enhance coverage during parallel sampling, we set the decoding temperature to 0.9 while retaining all other hyperparameters at their default values. All experiments are conducted on an NVIDIA A800 GPU with 80GB HBM3 memory to ensure a consistent runtime environment.

A.6 Detailed Insights

Insight 1: Task-specific preferences for model family and size combinations. We conduct preliminary tests across various combinations of model families and sizes on the MATH and MMLU datasets to explore the task-specific model preferences. Fig. 5(a–b) compares performance with various family combinations. The results show that allocating the budget to multiple instances of the strongest model is more effective than mixing families. For example, within the 3B space of LLaMA and Gemma on MMLU, LLaMA outperforms Gemma; thus, LLaMA$\times$2 surpasses both Gemma+LLaMA and Gemma$\times$2. This is because test-time scaling effectiveness is driven by the capability of base models, favoring replication of stronger ones. Fig. 5 (c–d) reports 10-run average performance with 90% confidence intervals under the same limited FLOPs budget, considering LLaMA 1B, 3B, and 8B, to explore whether limited budgets should be allocated to more small models or fewer large models (noting that with an unlimited budget, large models are always optimal). Reasoning tasks (MATH) favor mixtures of smaller models (3B$\times$3), while knowledge tasks (MMLU) prefer larger models (8B$\times$1). The trade-off depends on marginal performance gains: on MATH, LLaMA 3B improves by 7 points (from 39% to 46%) when scaled from one to two instances, showing the potential to surpass a single 8B (49%) with more instances, thus favoring small-model mixtures; on MMLU, the gain (41% to 45%) is modest, making larger models (8B$\times$1 with 64%) preferred. These are attributed to (i) task demands: reasoning tasks benefit from smaller-models ensembles because multiple models provide more opportunities to refine the answers with multi-step reasoning, whereas knowledge tasks need broad parametric knowledge coverage, better supported by large models; and (ii) task difficulty: easier tasks yield larger gains from small models, as they can already solve such tasks well and scaling further improves performance, whereas harder tasks are challenging for small models and demand large ones. Consequently, **tasks favor replication of the strongest model family, with small-model ensembles preferred only when their incremental gains are substantial.**

Insight 2: Parallel and sequential scaling saturate and decline beyond an optimal budget. Fig. 6 (a–b) shows that both parallel and sequential scaling on various datasets follow a non-monotonic pattern. Increasing the number of parallel nodes (width) or sequential nodes (depth) initially improves performance, but beyond a task-specific optimal point, performance plateaus and eventually declines. For example, peak performance is achieved at 8 parallel nodes or 8 sequential nodes on MATH, after which additional nodes yield no consistent gains. This performance degradation arises from different sources. In parallel scaling, performance converges once a sufficient width ensures dominance of correct answers, so additional nodes provide little benefit. Excessive outputs from preceding nodes lengthen input contexts, straining long-context capacity and degrading performance. In sequential scaling, performance improves while refinement benefits exceed potential propagated errors; once the refinement capacity is reached, additional steps mainly propagate and amplify errors, leading to performance degradation. In summary, **both width and depth exhibit task-dependent optima, beyond which extra computation provides negative returns.**

Insight 3: Interdependence between graph width and depth. Fig. 6 (c) shows MATH performance under varying width (parallel nodes) and depth (sequential nodes) combinations. We adopt a fixed architecture that first performs parallel sampling of w nodes from the input node, followed by sequential self-refinement of d nodes for each sampled branch, using the LLaMA-1B model uniformly across all nodes. To examine the trade-off between width and depth, we impose the constraint $wd \leq 24$. We observe: (i) accuracy at the optimal depth rises then falls as width increases (e.g., 38 at width 1, 47 at width 3, 45 at width 4), consistent with Insight 2; (ii) the optimal depth decreases with larger widths (e.g., 8 at width 1 vs. 4 at width 3), as initially wider structures enhance refinement capacity and accelerate convergence. Increasing depth yields the same pattern on width: accuracy follows a non-monotonic trend, and the optimal width decreases because deeper refinement allows correct answers to dominate earlier, shifting the optimal width point forward. In summary, **graph width and depth are interdependent, with growth in one dimension shifting the optimal point of the other.**

A.7 $f_{\text{cost}}(G, T)$ with the FLOPs compute metric

We adopt a simplified but standard FLOPs accounting scheme, where one multiply-add counts as 2 FLOPs, and causal self-attention reuses cached keys/values during decoding. Consider a model

at node v with non-embedding model parameters M, hidden size D, and layers L. Let N_p and N_d denote the input (prefill) and output (decode) lengths for node v on task $T = (\bar{N}_p^T, \bar{N}_d^T)$ where $\bar{N}_p^T$ and $\bar{N}_d^T$ are the average length of input and output, respectively.

Token-wise projection/MLP FLOPs. Each non-embedding weight is applied once per token through a matrix multiplication followed by addition, yielding approximately $2M$ FLOPs per token. Aggregating across sequence lengths, we obtain $2MN_p$ for prefill, $2MN_d$ for decode.

Attention FLOPs. For a single layer and a single head, the number of attention score dot-products (queries $\times$ keys) is:

- **Prefill (length N_p):** causal masking yields a triangular count

$$\sum_{i=1}^{N_p} i \;=\; \tfrac{N_p(N_p+1)}{2}.$$

- **Decode (length N_d):** token t attends to $N_p + t$ tokens, giving

$$\sum_{t=1}^{N_d} \big(N_p + t\big) \;=\; N_d N_p + \tfrac{N_d(N_d+1)}{2} \;=\; \tfrac{N_d(2N_p+N_d+1)}{2}.$$

Since each attention requires both query-key dot products and value applications, the total multiply-adds are $4LD$ FLOPs per token. Summing across D hidden size and L layers gives

$$\text{FLOPs}_{\text{attn, prefill}} = 2LD\, N_p(N_p+1), \qquad \text{FLOPs}_{\text{attn, decode}} = 2LD\, N_d(2N_p + N_d + 1).$$

These formulas combine constant factors from scoring, softmax, and value multiplication, while preserving quadratic and linear scaling in N_p and N_d.

Node-level cost. Summing the projection/MLP and attention costs yields

$$f_{\text{cost_prefill}}(N_p, M) \;=\; 2MN_p + 2LD\, N_p(N_p+1),$$
$$f_{\text{cost_decode}}(N_p, N_d, M) \;=\; 2MN_d + 2LD\, N_d(2N_p + N_d + 1),$$

so that

$$f_{\text{cost}}(N_p, N_d, M) = f_{\text{cost_prefill}}(N_p, M) + f_{\text{cost_decode}}(N_p, N_d, M).$$

Effective input length in a collaboration graph. In a multi-LLM collaboration graph $G = (\mathcal{V}, \mathcal{E}, \mathbf{R}, \mathbf{M})$, the effective prefill length for node v_i depends on the task average input and the number of predecessor outputs concatenated to its input. With $T = (\bar{N}_p^T, \bar{N}_d^T)$ and in-degree $d(v_i)$, we set

$$N_p^{v_i} = \bar{N}_p^T + d(v_i)\, \bar{N}_d^T, \qquad N_d^{v_i} = \bar{N}_d^T.$$

Graph-level cost. Summing node costs across the graph,

$$f_{\text{cost}}(G, T) = \sum_{v_i \in \mathcal{V}} \Big[f_{\text{cost_prefill}}\big(N_p^{v_i}, M_i\big) + f_{\text{cost_decode}}\big(N_p^{v_i}, N_d^{v_i}, M_i\big) \Big].$$

Substituting node-level cost formulas,

$$f_{\text{cost}}(G, T) = \sum_{v_i \in \mathcal{V}} \Big[2M_i N_p^{v_i} + 2L_i D_i\, N_p^{v_i}(N_p^{v_i}+1) + 2M_i N_d^{v_i} + 2L_i D_i\, N_d^{v_i}\big(2N_p^{v_i} + N_d^{v_i} + 1\big) \Big].$$

Simplified form. Let $A = \bar{N}_p^T$, $B = \bar{N}_d^T$, and $d_i = d(v_i)$. Then

$$f_{\text{cost}}(G, T) = \sum_{v_i \in \mathcal{V}} \Big[2M_i(A + d_i B) + 2L_i D_i (A + d_i B)(A + d_i B + 1) + 2M_i B + 2L_i D_i B\big(2(A + d_i B) + B + 1\big) \Big].$$

Expanding and grouping by d_i yields a quadratic form

$$f_{\text{cost}}(G, T) = \sum_{v_i \in \mathcal{V}} \big[\alpha_i\, d_i^2 + \beta_i\, d_i + \gamma_i \big],$$

with coefficients

$$\alpha_i = 2L_iD_iB^2, \quad \beta_i = 2M_iB+2L_iD_iB(2A+2B+1), \quad \gamma_i = 2(M_i+L_iD_i)(A+B)+2L_iD_i(A+B)^2.$$

Please remark that

(i) **Verifier/top-k filtering.** If a fuser applies top-k selection on predecessor outputs, replace $d(v_i)$ by $\min\{d(v_i), k\}$ in $N_p^{v_i}$.

(ii) **Alternative metrics.** For monetary cost, replace FLOPs-based node terms with calibrated surrogates $\{f_{\text{cost_prefill}}, f_{\text{cost_decode}}\}$ per model; graph aggregation remains identical.

(iii) **Budget normalization.** With unit budget defined as one inference of the smallest model, the normalized budget is

$$B = f_{\text{budget}}(G, T), \qquad f_{\text{cost}}(G, T) = B \cdot f_{\text{cost}}(G_{\text{smallest}}, T).$$

A.8 Detailed Budget Definition

Different model sizes and graph topologies incur substantially different computational costs: larger models introduce higher inference overhead, while denser topologies require more interactions. These differences make it challenging to establish a unified metric for budget measurement. To address this, we propose a *standardized budget definition* that enables comparability across model scales and topology complexities. For example, this framework allows us to equate the budget cost of "more sequential/parallel nodes with smaller models" to that of "fewer nodes with larger models."

Formally, let the average input and output lengths of a task be denoted by $T = (\bar{N}_p^T, \bar{N}_d^T)$. The total computational cost of a collaboration graph G on task T is defined as $f_{\text{cost}}(G, T)$, and the corresponding normalized budget is $B = f_{\text{budget}}(G, T)$. The cost function f_{cost} can be instantiated according to user preference to reflect different measures, such as FLOPs, wall-clock runtime, or monetary cost. To establish a common unit of comparison, we define the budget of executing one full inference with the smallest model in the pool as a single unit, i.e.,

$$f_{\text{budget}}(G_{\text{smallest}}, T) = 1,$$

where G_{smallest} denotes a graph consisting of only one node of the smallest model. Consequently, the budget value of any graph G is equivalent to the number of unit costs required, namely,

$$B = \frac{f_{\text{cost}}(G, T)}{f_{\text{cost}}(G_{\text{smallest}}, T)}.$$

where it means TTS graph with budget B is equal to run B-time single-node inference.

We define the computation cost of a multi-LLM collaboration graph G on a task $T = (\bar{N}_p^T, \bar{N}_d^T)$ in terms of FLOPs, which we adopt as the primary cost metric in this work. The corresponding cost function is stated in the theory below. The proof is in Appendix A.7.

FLOPs Cost Function*: For each node $v_i \in G$, the cost depends on the model size and its effective input/output lengths, leading to a quadratic dependence on the node in-degree $d(v_i)$. Summing across all nodes, the total cost can be expressed as*

$$f_{cost}(G, T) \;=\; \sum_{v_i \in \mathcal{V}} \big[\, \alpha_i \, d(v_i)^2 \;+\; \beta_i \, d(v_i) \;+\; \gamma_i \,\big],$$

where coefficients $\alpha_i, \beta_i, \gamma_i$ capture the contributions of model dimension, depth, and average task input/output lengths. Detailed derivations of $\alpha_i, \beta_i, \gamma_i$ are provided in Appendix A.7.

Choice of budget settings: We choose budgets 18, 42, and 80 to represent low-, medium-, and high-compute regimes under our FLOPs-based budget definition. These three settings roughly correspond to the compute required by 1×8B, $\sim 3 \times 8$B, and $\sim 5 \times 8$B model inferences, respectively. As shown in Table 1 and Table 4, we evaluate all methods under budgets 18, 42, and 80, and further analyze finer-grained budget ranges on MATH. A clear trend emerges: our method continues to improve as the budget increases, whereas most baselines improve from 18 to 42 but then decline from 42 to 80. We believe this reflects the value of our proposed insights, which enable more effective use of additional budget; without them, existing methods are more prone to over-expansion or inefficient allocation at higher budgets.

A.9 Detailed Optimization with Joint Objective

Our optimization objective is not limited to single-performance criteria; in many cases, it is necessary to identify graph structures that satisfy composite objectives, such as achieving both low latency and high accuracy. To this end, the proposed **Agent-REINFORCE** framework incorporates diverse feedback mechanisms obtained from the `Environment` to accommodate different optimization goals. For instance, under the joint objective of low latency and high performance, we incorporate the inference time of each candidate graph as an additional feedback signal to the `Agent`. Moreover, we can explicitly provide the `Agent` with prior knowledge through instructions that describe the relationship between graph structures and latency, for example, that latency is more sensitive to the number of nodes and the graph width, thereby accelerating the search for composite-optimal graphs. All feedback, including inference time, is stored in the `Archive`, enabling the LLM to leverage historical information to assess the marginal effect of latency reduction on performance, and thus achieve a principled trade-off between efficiency and accuracy.

A.10 Detailed Dollar Cost-based Budget

Table 6 is the API cost information for each model from Together AI and Compare Ai Models. We do not convert it in the same manner as above, as the dollar serves as a natural unit of price. Note that LLaMA-3.2 1B, Gemma-3 1B, and Gemma-1.1 2B are not quoted in Together AI or Compare Ai Models; for convenient consistency in our comparison, we adopt estimated reference values of 0.02, 0.02, and 0.06, respectively, for these models.

Table 6: Inference costs per 1M tokens for models from Together AI and Compare Ai Models.

Model Name	**Parameters**	**Inference Cost (per 1M tokens)**
LLaMA-3.1 70B	70B	\$0.88
LLaMA-3.1 8B	8B	\$0.18
LLaMA-3.2 3B	3B	\$0.06
Gemma-1.1 7B	7B	\$*0.27*

A.11 Detailed REINFORCE Algorithm

A gradient-based algorithm can be employed to solve the optimization problem. Since the search space of collaboration graphs is prohibitively large, exhaustive enumeration of all possible configurations is infeasible. Instead, we parameterize the distribution of graphs as $\tilde{\mathbf{G}} = \mathbb{P}_{\theta,\pi,\psi}$, where θ encodes the probabilities of edge existence, π encodes the probabilities of role assignments, and ψ encodes the probabilities of model selections.

Given a budget B, we set the number of nodes n to the maximum number of smallest models that the budget can cover. A straightforward approach to defining a parameterized distribution over DAGs with fixed n nodes, edges, models, and roles is as follows. We introduce real-valued parameters: $\theta = [\theta_{ij}],\ p_\theta(\theta_{ij}) = \sigma(\theta_{ij})$ for edge probabilities; $\pi = [\pi_1, \pi_2, \ldots, \pi_n]$ with role probabilities $p_\pi(r_i) = \mathrm{softmax}(\pi_i)$; and $\psi = [\psi_1, \psi_2, \ldots, \psi_n]$ with model probabilities $p_\psi(m_i) = \mathrm{softmax}(\psi_i)$. By iteratively refining this distribution, the algorithm progressively biases sampling toward low-loss collaboration graphs.

During training, we adopt the REINFORCE algorithm [57], a classical policy-gradient method that provides unbiased estimates of the utility gradient. It follows a sampling–gradient–update pipeline: candidates are sampled from the distribution, gradients are computed by evaluating on the training set, and parameters are updated via gradient ascent.

Monte Carlo Sampling. The probability of sampling a graph $G \sim \mathbb{P}_{\theta,\pi,\psi}$ is decomposed as

$$\mathbb{P}_{\theta,\pi,\psi} \;=\; p(\psi)\cdot p(\theta \mid \psi)\cdot p(\pi \mid \theta,\psi) \;=\; p(\psi)\cdot p(\theta \mid \psi)\cdot p(\pi \mid \theta),$$

where

$$p(\psi) = \prod_{i=1}^{n} p_\psi(\psi_i),$$

$$p(\theta \mid \psi) = \begin{cases} \prod_{i,j} p_\theta(\theta_{ij}), & \text{if the resulting graph is a DAG and } f_{\text{budget}}(G,T) \le B, \\ 0, & \text{otherwise,} \end{cases},$$

Algorithm 3 REINFORCE: Optimization of the Task-Specific Multi-LLM Collaboration Graph

Require: **Input:** task T, training data D_{train}, budget B, learning rate ℓ, batch size N
Ensure: optimized distribution $\mathbb{P}_{\theta,\pi,\psi}$ and final graph G^*

Initialize parameters θ (edge logits), π (role logits), ψ (model logits)
Define distributions: $p_\theta(e_{ij}) = \sigma(\theta_{ij})$, $p_\pi(r_i) = \mathrm{softmax}(\pi_i)$, $p_\psi(M_i) = \mathrm{softmax}(\psi_i)$
while stopping criterion is not met **do**
  $\mathcal{B} \leftarrow \emptyset$ (initialize mini-batch of sampled graphs)
  for $i = 1$ **to** N **do**
    Sample $G^{(i)} \sim \mathbb{P}_{\theta,\pi,\psi}$ (sample edges, roles, and models)
    if $f_{\text{budget}}(G^{(i)}, T) \le B$ **then**
      $u_i \leftarrow u_T(G^{(i)}, D_{\text{train}})$ (evaluate utility)
      $\mathcal{B} \leftarrow \mathcal{B} \cup \{(G^{(i)}, u_i)\}$
    end if
  end for
  $g_\theta \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(G,u)\in\mathcal{B}} u \nabla_\theta \log p_\theta(G)$
  $g_\pi \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(G,u)\in\mathcal{B}} u \nabla_\pi \log p_\pi(G)$
  $g_\psi \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(G,u)\in\mathcal{B}} u \nabla_\psi \log p_\psi(G)$
  $\theta \leftarrow \theta + \ell\, g_\theta$; $\pi \leftarrow \pi + \ell\, g_\pi$; $\psi \leftarrow \psi + \ell\, g_\psi$ (gradient ascent updates)
end while
Construct G^* by MAP decoding: include edge e_{ij} if $p_\theta(e_{ij}) \ge \tau_e$; set role $r_i \leftarrow \arg\max_r p_\pi(r_i{=}r)$; set model $M_i \leftarrow \arg\max_m p_\psi(M_i{=}m)$ (deterministic final graph)
Ensure $f_{\text{budget}}(G^*, T) \le B$ (greedy prune if needed)
return $\mathbb{P}_{\theta,\pi,\psi}$ and G^*

$$p(\pi \mid \theta) = \prod_{i=1}^{n} p_\pi(\alpha^{|d(v_i)|}\pi_i), \quad \alpha \in [1,\, 1.1],$$

where α is a constant that encourages the fusion role when the in-degree of v_i is high. This formulation provides a principled probabilistic parameterization of collaboration graphs, enabling efficient sampling and optimization within the REINFORCE framework.

Gradient Estimation. The gradient is calculated by:

$$\begin{aligned}\nabla_{\theta,\pi,\psi}\, \mathbb{E}_{G'\sim\mathbb{P}_{\theta,\pi,\psi}}\big[u_T(G')\big] &= \mathbb{E}_{G'\sim\mathbb{P}_{\theta,\pi,\psi}}\big[u_T(G')\, \nabla_{\theta,\pi,\psi} \log p_{\theta,\pi,\psi}(G')\big] \\ &\approx \frac{1}{N}\sum_{i=1}^{N} u_T(G^{(i)})\, \nabla_{\theta,\pi,\psi} \log p_{\theta,\pi,\psi}(G^{(i)}),\end{aligned} \tag{5}$$

where $G^{(i)}$ is the i-th candidate graph independently sampled from $\mathbb{P}_{\theta,\pi,\psi}$, and N is the number of Monte Carlo samples.

Parameter Updates. The distribution parameters are then updated with gradient ascent:

$$\begin{aligned}\theta &\leftarrow \theta + \tfrac{\ell}{N}\sum_{i=1}^{N} u_T(G^{(i)}) \nabla_\theta \log p_\theta(G^{(i)}), \\ \pi &\leftarrow \pi + \tfrac{\ell}{N}\sum_{i=1}^{N} u_T(G^{(i)}) \nabla_\pi \log p_\pi(G^{(i)}), \\ \psi &\leftarrow \psi + \tfrac{\ell}{N}\sum_{i=1}^{N} u_T(G^{(i)}) \nabla_\psi \log p_\psi(G^{(i)}),\end{aligned} \tag{6}$$

where ℓ is the learning rate.

Optimization loop. REINFORCE alternates between three phases: (i) *sampling*, where candidate graphs $G^{(i)}$ are drawn from the current distribution; (ii) *evaluation*, where utilities $u_T(G^{(i)})$ are

computed on the training set; and (iii) *update*, where parameters θ, π, ψ are refined by gradient ascent. This process repeats until convergence or when the budget is exhausted.

Final graph selection. After optimization, the learned distribution $\mathbb{P}_{\theta,\pi,\psi}$ is used to construct a deterministic collaboration graph G^*. Specifically, we decode by maximum a posteriori (MAP): edges are included if $p_\theta(e_{ij}) \geq \tau_e$, roles are assigned as $r_i = \arg\max_r p_\pi(r_i{=}r)$, and models are chosen as $M_i = \arg\max_m p_\psi(M_i{=}m)$. The final graph is pruned if necessary to ensure $f_{\text{budget}}(G^*, T) \leq B$. The complete optimization procedure is summarized in Algorithm 3.

A.12 Prompt Design in Agent-REINFORCE

We design structured prompts to guide the LLM search agent in initializing and updating the collaboration graph. Each prompt provides task context, distilled insights, and design constraints to support systematic reasoning and planning. For model family and size initialization, the agent ranks candidate families and sizes under budget constraints, guided by single-model performance and preliminary evaluations. This establishes a principled starting point for subsequent exploration. For model instance count initialization, the agent specifies concrete model combinations with family, size, and instance counts. These candidates are then tested in the environment, and the feedback highlights the most promising allocations. For graph updates, the agent leverages Insight 2, Insight 3, and feedback from the previous round to refine edge distributions, adjust connectivity, and balance budget allocation, thereby improving the overall structure and moving toward compute-optimal performance.

LLM Prompt for Model Family and Size Preference Initialization $\mathcal{A}$.`select_family_size()`

Your current task is model family and size initialization: you must provide the model family and size preferences for a test-time collaboration graph that will later be optimized into a DAG. An edge indicates that the previous model's output is the next agent's input.

========== TASK ==========
1. Examine the candidate model combinations listed at the end of this message.
2. Return a JSON dictionary of model family and size ranking.
No extra text, explanations, or formatting—just the dictionary.

========== INSIGHTS ==========
(1) Different tasks exhibit a clear preference for specific model combinations. Under budget constraints, it is necessary to identify the preferred model family and model size for each task.

========== DATA ==========
Single-model accuracy on `{task}` (higher is better):
`{model_profile}` or `{pre_test_accuracy}`
Random-graph pre-experiment results (including small models running once or twice and large models running once):
`{combinations_accuracy}`

========== CANDIDATES ==========
Choose only one from this list (each already fits the budget):
`{model_combinations}`

==========
Respond with the dictionary only. Example format:

LLM Prompt for Model Instance Counts Initialization $\mathcal{A}$.`select_instance()`

Your current task is model instance count initialization: you must provide the model instances for a test-time collaboration graph that will later be optimized into a DAG. We will test them in the Environment and select the best one according to the feedback. An edge indicates that the previous model's output is the next agent's input.

========== TASK ==========
1. Examine the model family and size preferences listed at the end of this message.

2. Return a JSON dictionary of model combinations with model family, size, and instance counts. No extra text, explanations, or formatting—just the dictionary.

========== INSIGHTS ==========
(1) Different tasks exhibit a clear preference for specific model combinations. Under budget constraints, it is necessary to identify the preferred model family and size for each task.

========== PREFERENCE ==========
Model family and size preferences:
`{model_family_size_preference}`

========== DATA ==========
Single-model accuracy on `{task}` (higher is better):
`{model_profile}` or `{pre_test_accuracy}`

==========
Respond with the dictionary only. Example format:

LLM Prompt for Graph Updates $\mathcal{A}$.`update()`

You are a professional Multi-LLM system optimizer. Your task is an iterative self-RL refinement of a multi-LLM system that solves the `{task}` dataset.

TASK CONTEXT
• A Multi-LLM system is represented as a directed acyclic graph (DAG). Each node = one language-model agent. Each directed edge = "the source agent's output is appended to the destination agent's context".
• For the current budget, we have a fixed model-selection requirement: `{model_selection}`
• You will see the last-round graph, its batch accuracy, and the full table of edge-selection probabilities.
• Your job: propose the next-round graph and the updated probability table, applying RL-style probability nudges.
• The graph you receive in this iteration has been expanded outward from the FinalDecision node, gradually increasing in both depth and breadth. The edge probabilities start with all edge probabilities set to zero, and through multiple sampling rounds, probabilities are raised only for edges that prove useful.

HISTORICAL SNAPSHOT
Last-round accuracy (`{task}`-dev batch): `{accuracy}`
Last-round graph: `{prev_graph}`
Last-round edge-probabilities: `{edge_probs}`

OPTIMIZATION RULES
R-1 Model counts must exactly match model selection after you assign models to all nodes.
R-2 A node's role is either "assistant" (generates a new answer) or "fuser" (reviews & picks the best).
R-3 Increase an edge probability only if it was sampled in the last-round graph AND proved useful. Always start expansion from FinalDecision's incoming edges, then its parents' incoming edges, and so on. Increase edges used by high-accuracy graphs, decrease edges from poor graphs.
R-4 Keep the graph acyclic; avoid too much in-degree to prevent context explosion; avoid very deep chains to prevent "answer corruption".

DATA AND INSIGHT
• Model accuracy on `{task}` (single-agent): `{model_profile}`
• The optimal depth is conditioned by current width, and vice-versa: wider graphs shift the depth sweet-spot downward, while deeper graphs reduce the optimal width.
• You should expand the architecture outward from the FinalDecision node, gradually adding depth and width.
• Different tasks favor different graph topologies; optimize toward the topology style that this task prefers.

WHAT TO RETURN
• graph — the next-round DAG, same schema as last-round graph.
• edge probs — the updated probability table, same schema and order as last-round edge-probabilities.

Example output format (do NOT add comments):
Graph: `{graph_example}`
Edge-probabilities: `{node_example}`

Now think step-by-step with the rules and insights above, and return the Graph and Edge-probabilities two blocks only.

A.13 Baselines

We compare three baseline categories: LLM-based (MaaO [19] and TextGrad [66]), gradient-based (GPTSwarm [77]), and traditional methods (Bayesian Optimization [48] and Random Search). Then, we detail their adaptation.

TextGrad [66] performs automatic "differentiation" through text, where an LLM generates a natural language "gradient" that guides updates to optimizable variables based on predictions and loss values. In the context of compute-optimal collaboration graph optimization for TTS, the probabilistic graph serves as the optimizable variable. Candidate graphs are sampled from the current distribution and evaluated on a batch of training data to compute the loss; the LLM then provides textual guidelines indicating how the graph should be refined given the observed loss and inputs. This process is repeated iteratively until convergence or a predefined stopping criterion is met. During initialization, TextGrad selects the maximal model combination that encompasses all potential candidates (i.e., allocating nodes to every feasible mixture of available models within the budget). Compared with our method, TextGrad lacks task-specific initialization and test-time scaling knowledge, making it a less efficient and less effective baseline.

MaaO [19] is a hybrid approach that integrates gradient-based optimization with LLM-guided optimization, leveraging the complementary strengths of both. Gradient-based methods provide precise directional updates in the parameter space but are prone to local optima, while LLM optimizers offer high-level heuristic guidance yet often lack stability. To address this, MaaO alternates between the two optimization strategies. In our problem setting of optimizing probabilistic graphs, we adopt REINFORCE to compute numerical gradients and use an LLM to generate textual updates, alternating between them during training. Concretely, the probabilistic graph is first initialized with a uniform distribution (same as described above), from which candidate graphs are sampled and evaluated on a training batch to compute predictions and loss values. Gradients derived from the loss are then used to update the probabilistic graph (see Appendix A.11). Subsequently, new candidates are sampled, and their losses are used by the LLM to provide textual updates on how the graph should be modified. This alternating process of gradient updates and LLM guidance continues until convergence or a stopping criterion is met.

GPTSwarm [77] generalizes LLM-based agent architecture search into a computational graph and optimizes it using gradient-based REINFORCE. In our problem setting, we adapt this approach as follows: a probabilistic graph is first initialized, from which candidate graphs are sampled and evaluated on a batch of training data to compute predictions and loss values. The loss gradients are then used to update the probabilistic graph, and this process is iterated until a stopping criterion is reached. The detailed REINFORCE optimization procedure is in Appendix A.11. However, as a purely gradient-based approach, GPTSwarm is relatively inefficient, as each update makes only incremental progress, and the method is susceptible to convergence at suboptimal local minima, thereby limiting both convergence speed and global search capability.

Bayesian Optimization (BO) [48] is a model-based framework for black-box optimization and has been widely applied to hyperparameter tuning. For optimizing collaboration graphs in test-time scaling, the graph is parameterized by θ, π, ψ, from which a concrete graph G is sampled and evaluated on a training batch to obtain its performance $f(G)$. Accordingly, BO treats θ, π, ψ as input variables, with the objective function defined as $F(\theta, \pi, \psi) = \mathbb{E}_{G \sim \mathcal{P}_{\theta,\pi,\psi}}[f(G)]$. Specifically, BO constructs a surrogate model, such as a Gaussian process, to approximate $F(\theta, \pi, \psi)$, and employs an acquisition function (e.g., Expected Improvement, EI) to guide the selection of promising candidates.

Each selected (θ, π, ψ) is evaluated by sampling multiple graphs to estimate average performance. Under budget constraints, the cost function $f_{\text{budget}}(G)$ can be incorporated via constrained acquisition (e.g., constrained EI). This iterative process of surrogate modeling, candidate selection, and evaluation continues until a stopping criterion is reached, at which point BO returns the optimal parameter set $(\theta^\star, \pi^\star, \psi^\star)$ and its corresponding high-performing probabilistic graph.

Random Search is a simple but widely adopted baseline in hyperparameter optimization. For compute-optimal collaboration graph search in test-time scaling, it generates candidate graphs uniformly at random under the budget constraint, without leveraging prior knowledge or performance history. While its simplicity makes it robust to irregular or non-smooth search landscapes and occasionally capable of identifying strong candidates, the absence of guidance typically leads to inferior search efficiency and performance compared with more structured or informed methods.

A.14 Convergence and Efficiency on MATH Dataset

As shown in Fig. 9, our method achieves the best accuracy and fastest convergence via strong initialization and guided by empirical insights. TextGrad tends to overuse the budget and slows down, while GPTSwarm/MaaO converges quickly but gets stuck in local optima.

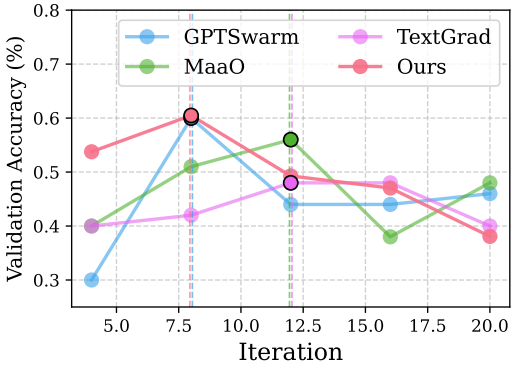

Figure 9: Training trajectories on MATH across LLM-based methods over 20 iterations. X-axis: iteration; Y-axis: the validation accuracy.

A.15 Detailed Agent Cost

To quantify the inference cost of the LLM agent during the search process, We further report the additional cost introduced by the LLM agent itself on the MATH dataset under a FLOPs budget of 80. Specifically, we separate the cost into the initialization stage and the iterative search stage, as shown in Table 7.

In terms of time cost, the search stage involves up to 30 iterations, corresponding to 30 batches of candidate trials. The total search time consists of two parts: trial generation by the LLM agent and trial execution in the Environment. Trial generation dominates the total time, accounting for $526.71/900.31 \approx 58.49\%$ of the total cost. Environment execution contributes $279.37/900.31 \approx 31.03\%$, while the initialization stage accounts for only $94.23/900.31 \approx 10.46\%$. In terms of dollar cost, the initialization stage only requires the LLM agent to propose suitable model combinations, resulting in a small cost of approximately 10.08% of the total. The remaining cost is incurred during the iterative search stage.

A.16 Additional Ablation Study

We conduct additional ablation studies by removing the initialization stage and the update stage, respectively. The pruning operation is required to satisfy the budget constraint, and therefore we do not ablate it. As shown in Table 8, both initialization and update are crucial to the effectiveness of Agent-REINFORCE. Removing either component leads to an approximately 10% performance drop.

Table 7: Agent cost on the MATH dataset under a FLOPs budget of 80.

Field	Value	Percentage
Agent total time (s)	900.31	100.00%
Initialization time (s)	94.23	10.46%
Trial generation time (s)	526.71	58.49%
Environment execution time (s)	279.37	31.03%
Total dollar cost (USD)	$0.12	100.00%
Initialization dollar cost (USD)	$0.01	10.08%
Search dollar cost (USD)	$0.11	89.92%

Table 8: Additional ablation study of Agent-REINFORCE.

Method	MATH Acc.	MMLU Acc.
Ours	56	54
w/o initialization	45	42
w/o update	44	47

A.17 A Qualitative Example of TTS Graph and its Output

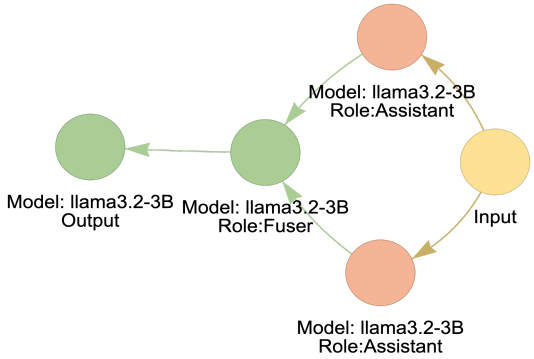

Figure 10: A qualitative example of TTS graph.

We present a qualitative example in Fig. 10 using an instance from the MATH dataset. The graph contains five nodes: one input node, one output node, two Assistant nodes, and one Fuser node. The topology consists of two Assistant nodes feeding into the Fuser node, which then connects to the output node. We provide the prompts and outputs for each node in the text boxes below. As shown, one Assistant node produces the correct answer while the other produces an incorrect one; the Fuser successfully selects the correct output. The output node receives input from a single node and therefore simply copies that node's output (details omitted for brevity).

Input Node

System Message: You are an expert mathematician. Help the user to solve this problem. Answer the following mathematics question. Provide your reasoning by showing your work before your answer. At the end of your response, output your final answer in the format: `'The answer is: [answer]'`. You must provide the separator `'The answer is: '` before your final answer. Make sure to put the answer (and only the answer) inside `\boxed{}`.

User Message: If $f(x) = \dfrac{a}{x+2}$, solve for the value of a so that $f(0) = f^{-1}(3a)$.

Assistant 1 Node Output

To find the inverse function $f^{-1}(x)$, we swap x and y in $f(x) = \frac{a}{x+2}$ and solve: $x = \frac{a}{y+2} \Rightarrow x(y+2) = a \Rightarrow xy + 2x = a \Rightarrow y = \frac{a-2x}{x}$, so $f^{-1}(x) = \frac{a-2x}{x}$.
We are given $f(0) = f^{-1}(3a)$. Compute $f(0) = \frac{a}{2}$. Next, $f^{-1}(3a) = \frac{a-6a}{3a} = -\frac{5}{3}$.
Set them equal: $\frac{a}{2} = -\frac{5}{3}$. Multiply by 6: $3a = -10$, so $a = -\frac{10}{3}$.
The answer is: `\boxed{`$-\frac{10}{3}$`}`.

Assistant 2 Node Output

To find the inverse function $f^{-1}(x)$, we swap x and y in $y = \frac{a}{x+2}$. After swapping, we get $x = \frac{a}{y+2}$, which simplifies to $xy + 2x = a$, then $xy = a - 2x$, and finally $y = \frac{a-2x}{x}$. Thus, $f^{-1}(x) = \frac{a-2x}{x}$.
We now solve for a such that $f(0) = f^{-1}(3a)$. Since $f(x) = \frac{a}{x+2}$, we have $f(0) = \frac{a}{2}$. Next, compute $f^{-1}(3a) = \frac{a-6a}{3a} = -\frac{5a}{3a} = -\frac{5}{3}$.
Setting them equal: $\frac{a}{2} = -\frac{5}{3}$. Cancelling $a \neq 0$ gives $\frac{1}{2} = -\frac{5}{3}$, which is impossible, so no real value of a satisfies the equation.

Fuser Node Prompt and Output

System Message:

You are a fusion agent in a multi-agent system. Your role is to review, condense, and synthesize the reasoning and answers produced by previous agents.

User Message:

According to the previous agents' answers for the problem:

Question: If $f(x) = \frac{a}{x+2}$, solve for a so that $f(0) = f^{-1}(3a)$.

As the fusion agent, synthesize a high-quality final answer by extracting correct reasoning, removing redundancy, ensuring mathematical rigor, and producing a clear justification. The final answer must be wrapped in `\boxed{}`.

Previous Answers:
1. Previous Agent (Version 1):

Correctly computed $f^{-1}(x) = \frac{a-2x}{x}$, $f(0) = \frac{a}{2}$, and $f^{-1}(3a) = -\frac{5}{3}$.
Correctly solved $\frac{a}{2} = -\frac{5}{3} \Rightarrow 3a = -10 \Rightarrow a = -\frac{10}{3}$.

2. Previous Agent (Version 2):

Derived $f^{-1}(x)$ by swapping x, y in $y = \frac{a}{x+2}$, leading to $x = \frac{a}{y+2}$, then $xy + 2x = a$, $xy = a - 2x$, and $y = \frac{a-2x}{x}$. Thus $f^{-1}(x) = \frac{a-2x}{x}$.
Computed $f(0) = \frac{a}{2}$ and $f^{-1}(3a) = \frac{a-6a}{3a} = -\frac{5}{3}$.
Equating gave $\frac{a}{2} = -\frac{5}{3}$, but incorrectly concluded no solution due to cancelling a incorrectly.

Fuser Output:

Synthesis of the Solution. Both previous agents correctly derived the inverse function $f^{-1}(x) = \frac{a-2x}{x}$. Agent 2 provided the cleanest evaluation of $f(0)$ and $f^{-1}(3a)$.
Step 1: $f(0) = \frac{a}{2}$.
Step 2: $f^{-1}(3a) = -\frac{5}{3}$.
Step 3: Set $\frac{a}{2} = -\frac{5}{3}$. Multiply both sides by 6 to clear denominators: $3a = -10$, so $a = -\frac{10}{3}$.
Step 4: Substitution verifies correctness.
Final Answer: `\boxed{`$-\frac{10}{3}$`}`.

A.18 Detailed Related Work

Test-time Scaling and Compute-optimal Strategy. Inspired by the human tendency to allocate additional cognitive effort for deeper and more deliberate reasoning, recent studies have proposed distributing extra computational resources during inference to improve model performance on various tasks [56, 55]. In parallel, other works [5, 58] have observed that increasing inference-time computation follows a scaling law analogous to that of training, where additional computation consistently enhances task performance. This phenomenon is commonly referred to as *Test-Time Scaling (TTS)*. Existing TTS techniques can be broadly categorized into two paradigms: *sequential scaling* and *parallel scaling*. In sequential scaling, the model enhances its reasoning ability by progressively extending a reasoning chain. A common approach is *self-refinement*, in which the model first generates an initial response and then iteratively revises it based on self-assessment [37, 15, 49, 11, 10]. Because this strategy depends heavily on the quality of the initial output, it tends to be more effective on relatively simple tasks [49]. By contrast, parallel scaling improves inference by generating multiple independent candidate solutions simultaneously and aggregating them into a final answer. Representative aggregation strategies include *majority voting* [35, 55], which selects the most frequent output among N candidates, and *Best-of-N* [5, 50, 17], which samples N solutions and uses a verifier to select the best one [47]. Other approaches employ LLMs themselves as fusers to integrate multiple candidates into a single output, thereby providing stronger generalization and flexibility [27, 30, 46]. Despite these successes, both paradigms exhibit limitations. Sequential scaling suffers from poor scalability, as extending the reasoning chain increases the risk of corrupting previously correct intermediate results [67]. Parallel scaling, while improving diversity, often lacks the depth of reasoning required

for more complex tasks [39]. To address these issues, hybrid approaches have been explored. For instance, Snell et al. [49] propose adaptively switching between sequential and parallel scaling depending on task difficulty, using sequential scaling for simpler tasks and parallel scaling for more complex ones. Other methods leverage tree-structured search to combine the two paradigms at the step or output level, employing process-level reward models to expand top-K intermediate steps and refine them further. Typical examples include beam search [64, 59] and Monte Carlo Tree Search (MCTS) [58, 49, 20, 52, 6, 72]. Nevertheless, most existing hybrid methods assume a *fixed inference structure* (e.g., fixed width or depth), limiting their flexibility. Recent studies have begun to relax these assumptions. For example, *Adaptive Parallel Reasoning* [43] dynamically switches between sequential and parallel computation using `spawn` and `join` operations, while *Adaptive Branching MCTS* unifies both paradigms within a tree-search framework, deciding at each node whether to parallelize candidate generation or continue sequential refinement. In addition, prior work has noted that sampling across multiple models naturally falls within the scope of test-time scaling, since ensembles improve diversity and output quality [71, 2, 27], yet this dimension remains underexplored in test-time scaling.

The configuration of allocating computation at inference time is central to the effectiveness of test-time scaling (TTS), giving rise to the *compute-optimal test-time scaling strategy*. A growing body of work [5, 58, 32, 65, 49, 53] highlights that model size and scaling configuration must be carefully balanced: in certain scenarios, smaller models can achieve superior accuracy compared to large models when constrained by the same compute budget. This line of research explores both model selection, deciding when to employ small versus large models, and method selection, choosing between alternative scaling paradigms to maximize utility. For instance, Snell et al. [49] show that the optimal scaling strategy varies with task difficulty: moderately challenging tasks favor parallel exploration with small models, whereas simpler tasks are better addressed through sequential refinement with large models. They further introduce a difficulty predictor to adaptively switch strategies. Other studies extend these ideas in different directions: Liu et al. [32] emphasize the sensitivity of scaling strategies to reward design, Yue et al. [65] develop a linear model to capture key determinants of scaling within retrieval-augmented generation (RAG), and Wu et al. [58] propose Reward Balanced Search (REBASE), a tree-search algorithm that achieves a Pareto-efficient balance between accuracy and inference cost through weighted voting. Despite these advances, existing approaches remain limited to fixed inference structures, overlooking the richer TTS patterns that arise in general graph topologies. Motivated by these gaps, we address a novel problem: unifying test-time scaling under a graph-based framework that incorporates heterogeneous model combinations, and searching for the compute-optimal collaboration graph.

Multi-agent Collaboration Graph. With the emergence of LLMs and the rapid development of LLM-based agents [12, 77], researchers have increasingly recognized that interactions among multiple agents can be naturally represented from a graph-based perspective [9, 77, 45, 36]. Graphs provide a principled abstraction for capturing communication patterns, role assignments, and coordination strategies in multi-agent systems, making them well-suited for reasoning about collaborative intelligence. Recent systems such as G-Designer [68], ARG-Designer [29], Heterogeneous Swarms [14], DyLAN [36], AgentNet [62], GPTSwarm [77], and MacNet [45] have explicitly employed graph structures to organize and optimize multi-agent interactions. These approaches primarily focus on structural optimization over a predefined set of agents, selecting the structure that maximizes task performance, which can be partially applied to our problem setting. However, they overlook the distinctive patterns of test-time scaling, resulting in inefficient architecture search.

LLMs for Optimization Optimization is fundamental to computational models and is often customized for individual tasks to address the challenges of complex decision spaces and performance landscapes. Large Language Models (LLMs), with their rich prior knowledge and reasoning capabilities, have opened new avenues for solving practical optimization problems [74, 19]. Existing research primarily employs LLMs in two paradigms: as black-box optimizers and in conjunction with gradient-based white-box optimization. The distinction lies in whether gradient information is available. In the black-box setting, LLMs are used to generate candidate solutions and iteratively refine them by leveraging their planning ability and extensive machine learning knowledge. Prior work has demonstrated the effectiveness of this approach in small-scale mathematical optimization [60, 73, 24], hyperparameter tuning [33, 34], and neural architecture search [76, 42, 26]. For instance, OPRO [60] proposed "optimization by prompting," where tasks are described in natural language and

LLMs iteratively generate new solutions based on meta-prompts and prior evaluations. AgentHPO [33] empowers LLMs to autonomously search hyperparameter configurations by processing task descriptions, conducting experiments, and refining search quality from accumulated trials. GENIUS [76] explored the potential of GPT-4 for neural architecture search, employing its generative ability as a black-box optimizer to efficiently navigate the search space and refine promising architectures. LLMs are particularly valuable during initialization, as they can generate high-quality solutions that embed prior knowledge, narrowing the search space and establishing a stronger foundation for subsequent iterations. This capability has also been applied to NAS initialization [25], genetic algorithms in bioengineering [41], and financial planning [13].

These studies demonstrate that LLMs can serve as general-purpose black-box optimizers. However, when gradient information is available—typically in data-rich scenarios—black-box optimization becomes inefficient, as each candidate must be evaluated on the full training set, leading to prohibitive search costs. To address this, recent work has combined gradient-based optimization with LLM-guided search to exploit their complementary strengths [19, 66]. For example, MaaO [19] interleaves gradient-based training with LLM-guided optimization, integrating the data efficiency and precise updates of gradient methods with the exploratory diversity of LLMs. TextGrad [66] generalizes this idea by transforming AI systems into computational graphs and using LLMs to generate textual updates that serve as a form of backpropagation. This framework provides natural language critiques of system components, such as neurons, prompts, molecules, or code segments, and guides their updates. Building on this line of work, we extend the complementary use of LLMs and gradient methods to compute-optimal test-time scaling by optimizing a gradient-available probabilistic graph. This approach enables us to combine the data efficiency of gradient-based optimization with the semantic task-awareness of LLMs, particularly for critical initialization and text-form parameter updates, thereby improving both search effectiveness and efficiency.

A.19 Limitations

Although some recent TTS methods operate at the token level by allocating additional compute during decoding [38, 40, 31, 61], this work focuses on solution-level TTS, where compute is allocated across complete candidate solution paths and collaboration structures under a fixed budget. While our method dynamically organizes heterogeneous LLMs, the effectiveness of multi-LLM collaboration is still bounded by the inherent capabilities of the underlying language models. Since different LLMs may produce slightly different outputs, the composition of the model pool can affect intermediate reasoning trajectories and ultimately influence the final prediction. Moreover, our current implementation focuses on text-based LLM collaboration, leaving the extension to multimodal information, such as images and audio, as an important direction for future work. Such extensions could provide richer contextual representations but may also introduce new challenges in graph construction, role assignment, and budget allocation. Finally, the discrete collaboration-graph search space grows rapidly with the budget, so the LLM-agent search overhead may increase substantially for larger budgets or larger model pools.

B LLM Usage

Large Language Models (LLMs) were used solely for language refinement, including rephrasing, grammar checking, and improving readability. They were not involved in ideation, methodology, experiments, or data analysis. All research concepts and results are the authors' own, and the authors take full responsibility for the manuscript, ensuring that LLM-assisted text complies with ethical standards and avoids plagiarism or misconduct.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions: formulating test-time scaling as a multi-LLM collaboration graph search problem, deriving empirical insights from pilot experiments, and proposing Agent-REINFORCE. The method is described as an LLM-agent-augmented framework that mirrors the REINFORCE pipeline, using feedback as a textual gradient to update the probabilistic graph and efficiently search for optimal multi-LLM collaboration graphs.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the Appendix A.19, including its focus on solution-level rather than token-level TTS, its current restriction to text-based LLM collaboration, and the dependence of multi-LLM collaboration performance on the composition and inherent capabilities of the model pool. It also notes that the discrete collaboration-graph search space grows rapidly with the budget, so LLM-agent search overhead may increase for larger budgets or larger model pools.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper contains a theoretical derivation of the FLOPs-based cost function for multi-LLM collaboration graphs. The assumptions and detailed derivation of the cost coefficients are provided in Appendix A.7, while the remaining contributions are primarily empirical and algorithmic rather than theoretical.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the experimental setup needed to reproduce the main results, including the evaluated datasets, model pool, baselines, budget settings, and convergence criterion. Additional implementation details, including budget definitions, prompt design, and cost-based evaluation, are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper uses publicly available benchmark datasets, including MATH, MMLU, and HumanEval, and publicly available model families. We provide the code and reproduction instructions in the supplemental material, including the implementation of Agent-REINFORCE, budget calculation, prompt design, and codes for reproducing the main experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the datasets and hyperparameters in the experiments. Additional details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars for the pilot insight analysis. For the main multi-LLM graph search experiments, the paper provides detailed results across multiple datasets, budget settings, baselines, and ablations to support the robustness of the main conclusions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix specifies the computer resources required to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is conducted in accordance with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential impacts of the work performed. It highlights positive impacts such as reducing unnecessary computation and deployment overhead through test-time scaling.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In the main text and appendix, all assets used in the paper are properly credited, and their licenses and terms of use are properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset introduced by the paper is the implementation of Agent-REINFORCE. We provide documentation alongside the released code.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted on publicly available benchmark datasets and existing model families.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: the paper explicitly describes the use of LLMs as a core component of Agent-REINFORCE. In particular, the LLM agent is used for model initialization, candidate graph sampling, and feedback-conditioned updates of the probabilistic graph, and the experiments specify DeepSeek-R1 as the LLM search agent.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.