

# COLT: Lightweight Multi-LLM Collaboration through Shared MCTS Reasoning for Model Compilation

Annabelle Sujun Tang<sup>1</sup> Christopher Priebe<sup>1</sup> Lianhui Qin<sup>1</sup> Hadi Esmaeilzadeh<sup>1</sup>

## Abstract

Model serving costs dominate AI systems, making compiler optimization essential for scalable deployment. Recent works show that a large language model (LLM) can guide compiler search by reasoning over program structure and optimization history. However, using a single large model throughout the search is expensive, while smaller models are less reliable when used alone. Thus, this paper seeks to answer whether multi-LLM collaborative reasoning relying primarily on small LLMs can match or exceed the performance of a single large model. As such, we propose a lightweight collaborative multi-LLM framework, dubbed **COLT**, for compiler optimization that enables coordinated reasoning across multiple models within a single Monte Carlo tree search (MCTS) process. A key contribution is the use of a single shared MCTS tree as the collaboration substrate across LLMs, enabling the reuse of transformation prefixes and cross-model value propagation. Hence, we circumvent both heavy internal reasoning mechanisms and conventional agentic machinery that relies on external planners, multiple concurrent LLMs, databases, external memory/versioning of intermediate results, and controllers by simply endogenizing model selection within the lightweight MCTS optimization loop. Every iteration, the acting LLM proposes a joint action: (compiler transformation, model to be queried next). We also introduce a model-aware tree policy that biases search toward smaller models while preserving exploration, and a course-alteration mechanism that escalates to the largest model when the search exhibits persistent regressions attributable to smaller models.

\*

<sup>1</sup>University of California, San Diego. Correspondence to: Annabelle Sujun Tang <sujun@ucsd.edu>.

Preprint. February 3, 2026.

\*Code is available at <https://github.com/he-aclab/COLT>.

## 1. Introduction

The cost of model serving increasingly dominates the lifecycle of modern machine learning systems, making inference efficiency critical to scalable adoption and growth (Patel & Ahmad, 2023). As such, compiler optimization for model serving is a key factor in enabling the broad deployment of these services. This is because the compiler directly determines the runtime behavior of neural workloads on target hardware by selecting and ordering transformations such as scheduling, fusion, and layout changes (Li et al., 2020). However, identifying effective optimization sequences for modern models is challenging, as the search space is exponentially large and characterized by complex interactions between transformations and hardware-specific constraints.

Common rule-based approaches (Whitfield & Soffa, 1990; Lerner et al., 2002; Kulkarni et al., 2007; 2009) fall short of addressing such complexities as they often rely on hand-tuned heuristics that can overfit to a specific workload or hardware target. This shortcoming propelled an era of compilers that generally relied on stochastic search (e.g., evolutionary strategies, simulated annealing, random search) to navigate the combinatorial search space of neural compilations (Vasilache et al., 2018; Chen et al., 2018a;b; Zheng et al., 2020; Shao et al., 2022). These approaches, although effective, suffer from the inherent sample inefficiency of the underlying stochastic search that treats the optimization as a black box, mostly context-insensitive exploration.

Most recently, a new wave of research has shown that large language models (LLMs) can be utilized for compiler optimization (Cummins et al., 2023; 2025; Deng et al., 2025; Pan et al., 2025; Tang et al., 2025). Rather than relying on fixed heuristics or pre-learned policies, these methods use LLMs to condition transformation proposals on the program representation, the sequence of prior optimization decisions, and observed intermediate outcomes. When integrated with structured search procedures, this form of contextual reasoning enables more efficient exploration of large optimization spaces and improves sample efficiency. However, these compiler approaches rely on a single large LLM, as smaller models are less reliable and often lead to degraded optimization trajectories (Pan et al., 2025; Tang et al., 2025). This design choice leaves open the question of whether reasoning

capacity must be concentrated in a single large model.

As such, the central research question that this paper seeks to answer is whether collaborative reasoning that relies primarily on small LLMs can match or exceed the performance of a single large model in LLM-guided compiler optimization, without introducing the heavy machinery of agentic systems. Naïvely combining multiple models does not resolve this challenge. Fixed routing strategies fail to adapt to the evolving state of the optimization process, while agentic systems introduce external controllers, multiple concurrent LLMs, external memory/version of intermediate results that complicate integration with compiler optimization pipelines (Abou Ali et al., 2025). Addressing this problem requires a mechanism that allows multiple models to participate in optimization while preserving a single, coherent optimization process with minimal control overhead.

To this end, we propose **COLL**aborative **LLM** reasoning via a shared **T**ree (**COLT**), a lightweight collaborative multi-LLM framework for compiler optimization in which multiple LLMs participate in a single Monte Carlo tree search (MCTS) (Browne et al., 2012) process without relying on any internal reasoning procedures. Rather than employing external planners or controllers, model selection is *endogenized* within the optimization loop. Specifically, each node in the MCTS represents a joint state consisting of both the current program and the acting model. At each iteration, the selected LLM proposes (1) a compiler transformation to update the program state and (2) a recommendation for the next model to query, thereby expanding the tree along both dimensions. This design circumvents heavy internal reasoning mechanisms as well as conventional agentic machinery based on external planners, multiple concurrent LLMs, databases, external memory/versioning of intermediate results, and centralized controllers. Instead, all coordination emerges naturally from the lightweight MCTS optimization process through endogenous model selection. We further introduce a model-aware tree policy that biases search toward smaller LLMs while preserving sufficient exploration, along with a course-alteration mechanism that selectively escalates to the largest LLM when the search exhibits persistent regressions attributable to small-model proposals.

A key novel aspect of **COLT** is the use of a shared MCTS tree as the collaboration substrate across multiple LLMs, alleviating the need for launching multiple concurrent LLMs or using external memory and planning to track these LLMs. Moreover, rather than exploring disjoint optimization trajectories, all models operate over the same evolving search structure, allowing proposals from different LLMs to branch from and extend common partial transformation sequences. Because model-routing decisions are embedded directly in the tree, different models can contribute to and build upon a shared optimization history. Crucially, value esti-

mates obtained from downstream program variants, resulting from both model-selection decisions and transformation choices, are backpropagated through the shared tree. This shared backpropagation allows information discovered by one model to inform subsequent decisions made by others, which is particularly effective for compiler optimization settings where transformation sequences exhibit compounding effects and long-range interactions.

Across representative benchmarks drawn from modern model architectures and evaluated on both CPU and GPU platforms, our approach consistently exceeds the optimization quality of a single large-model baseline. Averaged across our five benchmarks, **COLT** achieves  $10.86\times$  speedup on CPU ( $30.05\times$  on GPU) for its best configuration and outperforms the single-model baseline while having only 23.9% of the total calls to the largest LLM. The results demonstrate the efficacy of our lightweight approach in realizing collaboration among LLMs through a shared tree.

## 2. Endogenous Model Selection for Multi-LLM Collaborative Compilation

### 2.1. Model Selection as a Sequential Decision Problem

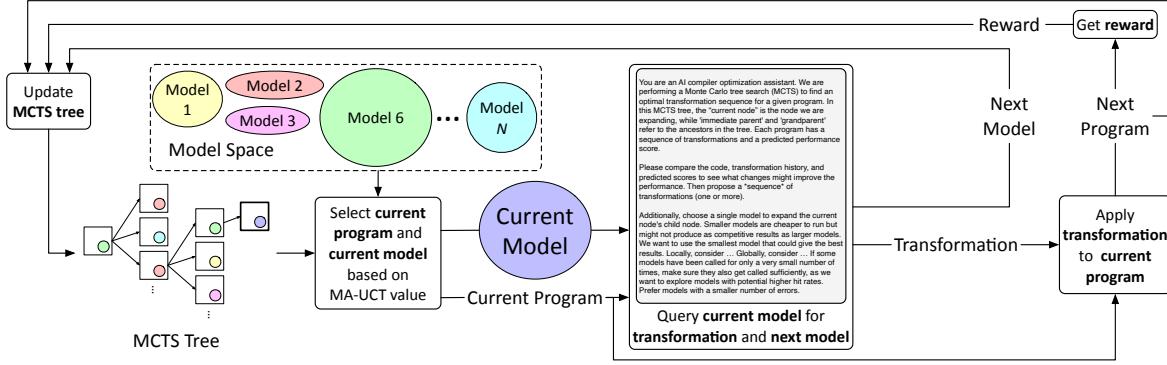
We consider the problem of selecting which LLM to invoke during an optimization process that unfolds over multiple steps. At each step, the choice of model affects not only the immediate quality of the proposed action but also the future trajectory of the optimization process. As a result, model selection cannot be treated as an independent or myopic decision. Instead, it constitutes a sequential decision problem in which early model choices influence downstream opportunities, failures, and recovery behavior.

Let  $M = \{m_1, \dots, m_N\}$  denote a finite set of available LLMs, which may differ in size, capability, and reliability. When invoked, a model produces a stochastic proposal conditioned on the current optimization context. Importantly, the consequences of invoking a particular model affect the future state of the optimization process and the decisions that follow. This coupling makes static routing rules or greedy selection strategies insufficient, as they cannot account for long-horizon effects.

Our goal is therefore to embed model selection directly into the optimization process. In doing so, decisions about which model to invoke can be evaluated under the same long-horizon objective as the actions those models propose.

### 2.2. Compiler Optimization as the Underlying Environment

We instantiate this sequential decision problem in the setting of compiler optimization for neural workloads. Following prior work, we consider the task of optimizing an initial program  $p_0 \in P$  through the repeated application



**Figure 1.** Overview of **COLT**. LLMs collaborate through the shared MCTS tree to propose reward-increasing transformations.

of semantic-preserving compiler transformations (Touati & Barthou, 2006). Each transformation  $o \in O$  maps one program to another. The objective is to identify a transformation sequence of length  $T$  that maximizes a performance metric  $f : P \rightarrow \mathbb{R}_{\geq 0}$  measured on a target hardware platform.

Following prior work, this process can be naturally modeled as a finite-horizon Markov decision process (MDP) (Kulkarni & Cavazos, 2012). States correspond to programs. Actions correspond to compiler transformations. Transitions are deterministic given a transformation. Rewards are derived from the performance metric.

Although the environmental dynamics are deterministic, the optimization procedure itself is stochastic. This stochasticity arises from LLM-driven proposal generation. This formulation captures the structure of compiler optimization, but does not yet account for the choice of which LLM generates proposals at each step.

### 2.3. Joint Decision Process with Endogenous Model Selection

To make model selection a first-class decision variable, we extend the compiler optimization formulation by coupling program evolution with model selection. Specifically, we define a joint state space  $\mathcal{S} = P \times M$ . Each state  $\langle p_t, m_t \rangle$  represents both the current program and the identity of the model responsible for proposing the next action. Actions in the joint process are pairs  $\langle o_t, m_{t+1} \rangle \in O \times M$ . Each action consists of a compiler transformation to apply and a recommendation for which model should be invoked next. When the current state is  $\langle p_t, m_t \rangle$ , the active model  $m_t$  induces a stochastic proposal distribution over such joint actions. Applying an action deterministically produces the next state  $\langle o_t(p_t), m_{t+1} \rangle$ .

This construction yields a unified decision process. Both transformation and model-selection decisions are evaluated under the same long-horizon objective. Model selection is not handled by external controllers or fixed policies. Instead, it is endogenized within the optimization process itself. As a result, the optimizer can reason about when smaller LLMs

are sufficient, when escalation is necessary, and how LLM choices interact with the evolving program state.

This formulation also induces a natural  $N$ -armed contextual bandit subproblem at each decision point. Conditioned on the current program state  $p_t$  and the current model  $m_t$ , selecting arm  $m_t$  yields a stochastic outcome consisting of a proposed joint action  $\langle o_t, m_{t+1} \rangle$  together with the resulting reward derived from applying  $o_t$  to  $p_t$ . Thus, each LLM defines a distinct state-conditional proposal distribution over joint actions, and the optimizer repeatedly faces the problem of choosing which LLM to invoke in the current context. However, because the objective depends on the quality of the entire transformation sequence rather than a single step, this bandit view is coupled to a long-horizon optimization problem. Consequently, the optimizer must simultaneously balance exploration and exploitation over LLMs while reasoning over long-horizon effects of transformation sequences within a single unified decision process.

## 3. A Lightweight Framework for Collaborative Multi-LLM Compilation

With this formulation, we present **COLT**, a framework for lightweight multi-LLM collaborative compiler optimization in which multiple models participate in a single Monte Carlo tree search (MCTS) (Browne et al., 2012) process. Rather than treating the choice of LLM as an external routing problem or a fixed design choice, **COLT** embeds model selection directly into the optimization process itself. At every step, the optimizer jointly decides what transformation to apply and which model should reason about the next decision.

In existing LLM-based compiler optimization frameworks (Cummins et al., 2023; 2025; Deng et al., 2025; Pan et al., 2025; Tang et al., 2025), decisions are concentrated in a single LLM. While effective, this approach obscures a critical degree of freedom: different LLMs exhibit distinct trade-offs, and invoking a particular model conditions the structure of subsequent decisions. An early reliance on a smaller LLM can induce compounding errors, while premature escalation to a larger LLM can waste capacity on

decisions that do not require it. We postulate that capturing these interactions requires reasoning about LLM choice under the same long-horizon objective as transformations.

**COLT** enables such reasoning without introducing the heavy machinery of agentic systems. There is no external controller, no hierarchical planner, no concurrent LLMs, and no long-lived agent state or versioning. Instead, collaboration emerges from a single, coherent optimization process in which multiple models participate by proposing actions within a shared search structure. To bias search toward efficient collaboration, we introduce a model-aware tree policy that favors smaller models while preserving exploration. In addition, we incorporate a lightweight course-alteration mechanism that selectively escalates to the largest LLM when the search exhibits persistent regressions. These mechanisms allow model selection decisions to be evaluated and reinforced through their downstream effects, enabling smaller LLMs to drive the majority of the optimization process without sacrificing robustness.

A key aspect of **COLT** is that collaboration among multiple LLMs is enabled by organizing all interaction around a single shared MCTS tree that jointly represents model-selection and compiler transformations decisions. Because all LLMs operate over the same evolving tree, proposals from different LLMs branch from and extend shared transformation prefixes, allowing partial optimization trajectories to be reused rather than explored independently. By making model choice part of the tree itself, the framework ensures that routing decisions and transformation decisions are evaluated under a unified long-horizon objective. Moreover, value estimates obtained from downstream program variants, arising from transformation choices and model-routing decisions made by different models, are backpropagated through one shared tree, allowing information discovered by one model to inform subsequent decisions made by others. Crucially, this shared search structure serves as the sole coordination mechanism among models, making effective collaboration possible without external controllers, explicit communication, or agentic state layered on top of the optimizer. This form of shared backpropagation across both transformations and model-selection decisions is what makes the proposed framework effective for compiler optimization, where transformation sequences exhibit compounding effects and long-range interactions.

### 3.1. Tree Search with Endogenous Model Selection

We instantiate the joint decision process introduced in §2.3 using a Monte Carlo tree search (MCTS) procedure in which LLM selection is treated as an explicit, first-class decision. Figure 1 illustrates the resulting optimization loop. Each node in the search tree corresponds to a joint state  $\langle p, m \rangle \in P \times M$ , consisting of a program state paired with the LLM responsible for generating the proposal. During

each iteration, tree expansion begins by selecting a node using the model-aware tree policy described in §3.2. The associated LLM  $m$  is then queried to propose a joint action  $\langle o, m' \rangle \in O \times M$ , which specifies both a candidate compiler transformation and the LLM to invoke at the next step. This design allows consequences of model-selection choices to be evaluated under the same long-horizon objective as transformation sequences themselves, as follows.

Applying the transformation produces a new program state, from which the search performs a rollout consisting of a short sequence of randomly selected transformations. The terminal program produced by the rollout is evaluated using a cost model, which efficiently estimates the objective and avoids the latency of direct execution on target hardware. In this work, we leverage TVM’s unmodified hardware-agnostic cost model (Chen et al., 2018a;b), which is based on XGBoost (Chen & Guestrin, 2016). The resulting scalar reward is then propagated back along the selected path, updating visit counts and value estimates at each node. Because rewards depend solely on the resulting program, credit assignment naturally spans both compiler transformations and the associated model-selection decisions, allowing their long-horizon effects to be reinforced or corrected within a single unified search tree.

### 3.2. Model-Aware Tree Policy

Node selection in MCTS is governed by a tree policy that balances exploration of less-visited areas with exploitation of high-rewarding regions. Endogenizing model selection requires the search policy to express preferences over *how* optimization trajectories are explored, without changing *what* constitutes an optimal solution. While compiler optimization continues to optimize solely for program performance, the tree policy must allocate exploration effort across trajectories that invoke models of differing sizes. Incorporating model size into the reward would conflate these concerns and allow inferior programs to be preferred due to smaller LLMs. Instead, model size should act as a prior over exploration rather than as part of the optimization objective.

We address this by introducing a *model-aware* variant of the Upper Confidence Bounds applied to Trees (UCT) selection rule referred to as MA-UCT. The resulting MA-UCT score augments the standard UCT (Kocsis & Szepesvári, 2006) formulation with an explicit, normalized preference for smaller models, while retaining a reward term based exclusively on program performance and a conventional exploration bonus. Specifically, for a node corresponding to program  $p_i$  and model  $m_i$ , we define the MA-UCT score as

$$\begin{aligned} \text{MA-UCT}(p_i, m_i) = & (1 - \lambda) \frac{W(p_i, m_i)}{N(p_i, m_i)} + \lambda \phi_{\text{small}}(m_i) \\ & + c \sqrt{\frac{\ln N(p_{i-1}, m_{i-1})}{N(p_i, m_i)}} \end{aligned}$$

where  $N(p_i, m_i)$  denotes the visit count of the node and  $W(p_i, m_i)$  is the cumulative reward of the corresponding program. The function  $\phi_{\text{small}}(m)$  encodes a normalized preference for smaller models and is defined as

$$\phi_{\text{small}}(m) = \frac{\log n_\theta(m_{\max}) - \log n_\theta(m)}{\log n_\theta(m_{\max}) - \log n_\theta(m_{\min}) + \varepsilon} \in [0, 1]$$

where  $n_\theta(m)$  denotes the parameter count of model  $m$ , and  $n_\theta(m_{\min})$  and  $n_\theta(m_{\max})$  are the minimum and maximum model sizes in the set of candidate models. The parameter  $\lambda \in [0, 1]$  controls the trade-off between optimizing program reward and preferring smaller models, while the exploration term preserves standard UCT behavior. This structure biases search toward trajectories in which smaller models are sufficient, without preventing the selection of larger models when justified by downstream rewards. As a result, model preference shapes the allocation of search effort rather than the definition of optimality, preserving the long-horizon objective of compiler optimization while improving the efficiency of collaborative multi-LLM search.

### 3.3. LLM-Guided Contextual Model Selection

Model selection is treated as an integral optimization decision, grounded in the current program state, recent optimization trajectory, and observed outcomes during search.

**Contextual prompt design.** At each node expansion, the active LLM is prompted with the current program, the sequence of transformations along the search path, and the candidate set of available LLMs. This prompt is augmented with both global and local statistics summarizing prior LLM behavior. Global statistics include, for each model, (1) a hit rate measuring the fraction of invocations that improve predicted reward, (2) the total number of invocations, and (3) an error count capturing invalid transformations or invalid model identifiers. Local context exposes the identities and parameter counts of LLMs used at the current node and its immediate ancestors, along with predicted reward differences between successive programs. Together, these signals provide a compact summary of model reliability, recent progress dynamics, and the surrounding search regime.

**Size-aware selection and constraints.** Given this combined context, the model proposes a single next model to invoke. The prompt emphasizes a size-aware objective: smaller models are preferred whenever they are likely to suffice, while larger models may be selected when higher capacity is needed. The goal is to identify the smallest model likely to support continued progress, rather than to trade off model size against program quality. To encourage robustness and exploration, models with limited prior usage are softly encouraged, the largest model is required to be invoked for a minimum fraction of total calls, and models with high error rates are discouraged. These preferences bias selection toward efficient and stable behavior without

imposing hard routing rules or external control logic.

Each LLM invocation produces a joint proposal consisting of a compiler transformation and a recommendation for the next model to invoke. This recommendation is treated as advisory rather than prescriptive: its influence on the optimization trajectory is mediated by MCTS and the model-aware tree policy (§3.2), allowing model-selection decisions to be evaluated over long horizons. As the search progresses, this formulation enables adaptive collaboration among LLMs. Smaller LLMs dominate routine decisions, while larger LLMs are selectively invoked for early exploration or navigation of unfamiliar regions of the optimization space.

### 3.4. Course Alteration via Large-Model Intervention

**COLT** is designed to rely primarily on smaller LLMs for efficiency, but prior work (Pan et al., 2025; Tang et al., 2025) and our own findings (see §4.2) show that when a single small model is used throughout the optimization process, small local errors can accumulate and degrade search quality. When used aggressively, such errors can cause the search to drift into locally degraded regions of the program space. We therefore introduce a *course alteration* mechanism that intermittently employs the largest LLM to reorient exploration when persistent degradation is detected.

**Diagnosing persistent small-model regressions.** Course alteration is evaluated only when the currently active LLM is a small one, defined as any model other than the largest in the candidate set. This asymmetry reflects a deliberate size-aware design: smaller LLMs are preferred for efficiency, while the largest LLM is reserved as a higher-capacity resource that can be invoked sparingly to alter the search trajectory. The mechanism inspects the current root-to-node path in the MCTS tree after a child has been expanded and rolled out, but before any backpropagation is performed. A regression is defined as a transformation that yields a program whose performance (evaluated via  $f$ ; see §2.2) is worse than that of its parent. Two regressions are considered consecutive if they occur along the same path and are both attributable to small-model invocations, regardless of whether they are adjacent in the tree. When such a pattern is detected, the regressive child expansion is pruned, preventing degraded trajectories from influencing value estimates.

**Large-model intervention as exploratory alteration.** Following pruning, control is temporarily escalated to the largest LLM in the candidate set. Rather than assuming a definitive correction, the intervention introduces a higher-capacity proposal at the same program state to alter the direction of exploration. Concretely, the parent node is re-expanded under the large model using the same optimization context, and backpropagation proceeds only with statistics from this altered continuation.

Overall, course alteration implements a capacity-aware but

non-oracular collaboration strategy. Smaller LLMs remain the primary drivers of search due to their efficiency, while the largest LLM is used sparingly as a high-capacity perturbation when progress stalls. Control returns immediately to the collaborative search process after intervention, and no model is assumed to be universally optimal.

## 4. Evaluation

### 4.1. Methodology

**Benchmarks.** We evaluate our framework on five representative computational kernels drawn from production-scale models: (1) a self-attention layer from Llama-3-8B ([Llama Team, AI @ Meta, 2024](#)), (2) a mixture-of-experts (MoE) layer from DeepSeek-R1 ([DeepSeek-AI, 2025](#)), (3) a self-attention layer from FLUX (stable diffusion) ([Black Forest Labs, 2024](#)), (4) a convolution layer from FLUX ([Black Forest Labs, 2024](#)), and (5) an MLP layer from Llama-4-Scout ([Meta, 2025](#)). In addition, we include an end-to-end evaluation of Llama-3-8B ([Llama Team, AI @ Meta, 2024](#)) to assess the impact of optimization decisions at the full model level. The target hardware platform of our main experiments is an Intel Core i9.

**Baselines.** We evaluate single-model baselines by fixing the LLM for all MCTS expansions and disabling model-selection decisions while keeping the search procedure otherwise identical. We consider a single-large-model configuration using GPT-5.2 ([OpenAI, 2025](#)) and a single-small-model configuration using gpt-5-mini ([OpenAI](#)).

**Implementation.** We implement our framework on top of TVM’s MetaSchedule ([Chen et al., 2018a; Shao et al., 2022](#)) by replacing its default search routine with MCTS whose state consists of a schedule trace over a fixed base IRModule and the identity of an LLM. Our MA-UCT (see §3.2) criterion is realized with  $\lambda = 0.5$ , exploration parameter  $c = \sqrt{2}$ , and branching factor  $B = 2$  following prior work ([Coulom, 2007; Auer et al., 2002](#)). When expanding a node, we materialize the current program from the trace and generate a prompt that includes (1) the current program/schedule, (2) a compact summary of recent search history (parent/grandparent traces and outcomes), (3) the allowable next transformations, and (4) search- and model-context (see §3.3). The queried LLM outputs both the next schedule edit (a valid transformation) and the identity of the next model. The resulting candidate is evaluated via TVM’s unmodified cost-model on the target hardware platform and incorporated into the MCTS tree with back-propagation. We inherit hardware agnostic capabilities from TVM’s cost-model, which is based on XGBoost ([Chen & Guestrin, 2016](#)) and can capture different hardware. Course alteration is handled via the tree manager.

**Experimental environment.** All experiments are conducted using Apache TVM v0.20.0 ([Chen et al., 2018a;](#)

[Apache TVM Community, 2025](#)). Additionally, we leverage OpenAI and Nscale model serving APIs to access the respective models. We repeat each experiment 10 times and report the mean performance to ensure statistical stability.

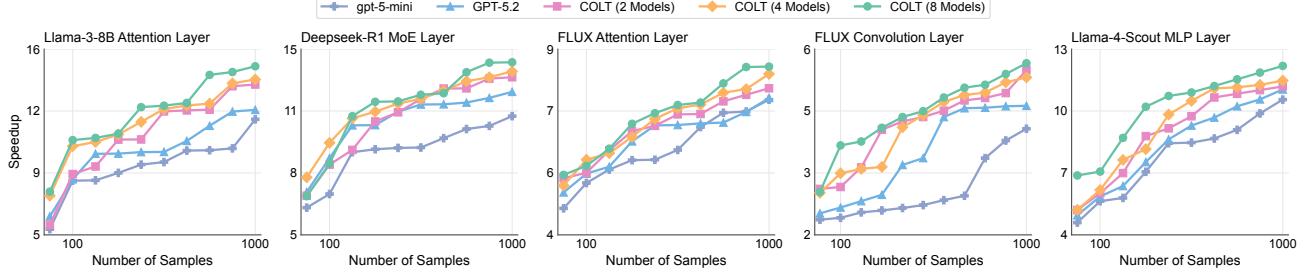
**Performance measurement.** We use execution latency as the primary metric and report speedup, defined as the latency of the original, unoptimized IRModule divided by the latency after applying the selected transformations. Latency is measured directly on the target hardware. Additionally, we report sample efficiency (the maximum speedup achieved divided by the number of samples evaluated) and sample efficiency gain (the ratio of sample efficiencies between experiments). We also report model invocation rate, which measures the fraction of total calls attributed to each model excluding course alterations over the entire search process, as well as the course alteration rate, measuring the fraction of course alterations relative to total small model calls.

### 4.2. Experimental Results

We present the impact of multi-LLM collaboration on optimization quality by analyzing how speedup improves with an increasing search budget. Figure 2 illustrates the speedups achieved across our five benchmarks for five configurations, including the single-model baselines and collaborative model sets of increasing size. Specifically, we consider: (1) **COLT(2 Models)**, a two-model set consisting of GPT-5.2 and gpt-5-mini; (2) **COLT(4 Models)**, a four-model set obtained by additionally including DeepSeek-R1-Distill-Qwen-32B ([DeepSeek-AI, 2025](#)) and Llama-3.1-8B-Instruct ([Llama Team, AI @ Meta, 2024](#)); and (3) **COLT(8 Models)**, an eight-model set further augmented with DeepSeek-R1-Distill-Qwen-7B ([DeepSeek-AI, 2025](#)), Qwen3-8B ([Team, 2025](#)), Qwen3-14B ([Team, 2025](#)), and Devstral-Small-2505 ([Mistral AI, 2025](#)).

**Single small vs. single large model.** Across all benchmarks in Figure 2, the single-small-model baseline (gpt-5-mini) consistently underperforms the single-large-model baseline (GPT-5.2), and in several cases achieves comparable performance only at best. This indicates that, when used in isolation, small models are generally less effective at proposing high-quality optimization sequences than a large model.

**Effect of collaboration.** Fixing the model to a single large LLM throughout the search yields lower final speedups than any **COLT** configuration. In Figure 2, on the Llama-3-8B Attention Layer, **COLT(8 Models)** attains  $14.98 \times$  final speedup, while GPT-5.2 only reaches  $12.40 \times$  final speedup, a 20.8% improvement. Additionally, collaboration often yields large early gains: on FLUX Convolution Layer at 100 samples, **COLT(8 Models)** reaches  $3.83 \times$  speedup, a 49.6% improvement over the single-large-model baseline (GPT-5.2). This suggests that collaboration enhances the quality of proposed transformations compared to a single model.


 Figure 2. **COLT** where the largest model is GPT-5.2.

**Table 1.** Invocation rates (%) of different LLMs in multi-LLM compilation across five benchmarks on CPU/GPU.

Largest Model: GPT-5.2 CPU/GPU				Largest Model: Llama-3.3-70B-Instruct			
Model	COLT (2 Models)	COLT (4 Models)	COLT (8 Models)	Model	COLT (2 Models)	COLT (4 Models)	COLT (8 Models)
GPT-5.2 (Regular)	27.4/29.3	20.2/18.3	12.3/11.3	Llama-3.3-70B-Instruct	49.4	34.0	21.2
gpt-5-mini	72.6/70.7	41.4/38	9.1/2	gpt-5-mini	50.6	31.0	3.4
DeepSeek-R1-Distill-Qwen-7B	—	—	14.5/29.3	DeepSeek-R1-Distill-Qwen-7B	—	—	10.2
Llama-3.1-8B-Instruct	—	17.2/33.3	13.1/8.6	Llama-3.1-8B-Instruct	—	25.6	17.8
Qwen3-8B	—	—	40.5/25.2	Qwen3-8B	—	—	25.0
Qwen3-14B	—	—	5.4/17.1	Qwen3-14B	—	—	17.3
Devstral-Small-2505	—	—	2.8/1.3	Devstral-Small-2505	—	—	4.0
DeepSeek-R1-Distill-Qwen-32B	—	21.2/10.4	2.3/5.2	DeepSeek-R1-Distill-Qwen-32B	—	9.4	1.1
Course Alteration Rate	16.2/16.7	17.1/16.9	17.4/17.3	Course Alteration Rate	16.9	17.0	17.5
GPT-5.2 (Total)	35.1/36.7	29.8/28.3	23.9/23.1	Llama-3.3-70B-Instruct (Total)	53.4	40.6	30.7

**Scaling the model set.** In Figure 2, among the **COLT** configurations, performance improves monotonically as the candidate model set is expanded, with **COLT(8 Models)** achieving the highest final speedups across all benchmarks. Notably, as shown in Table 1, for **COLT(8 Models)**, GPT-5.2 accounts for only 12.3% of regular model invocations (23.9 % including course alteration), while there is a shift of regular model invocations to mostly smaller open-source models, such as Qwen3-8B at 40.5% of calls, DeepSeek-R1-Distill-Qwen-7B (14.5%), and Llama-3.1-8B-Instruct (13.1%). This trend indicates that the benefits of collaboration scale with the size and diversity of the model set. Such behavior is consistent with prior evidence in machine learning that aggregating smaller predictors, e.g., via boosting (Ferreira & Figueiredo, 2012), can yield stronger overall performance than any individual model alone.

**End-to-end analysis.** Table 2 summarizes end-to-end Llama-3-8B tuning with realized samples, final speedup, and sample-efficiency gain relative to gpt-5-mini. Using GPT-5.2 as the largest model, **COLT(4 Models)** achieves 5.02× speedup with only 390 samples and has the strongest efficiency gain (1.55×), which matches the single-large baseline’s 5.01× speedup while requiring substantially fewer samples (530 for GPT-5.2). Overall, collaborative configurations show improved sample efficiency over single-model baselines on end-to-end workloads.

#### 4.3. Ablation Study: Largest Model Choice

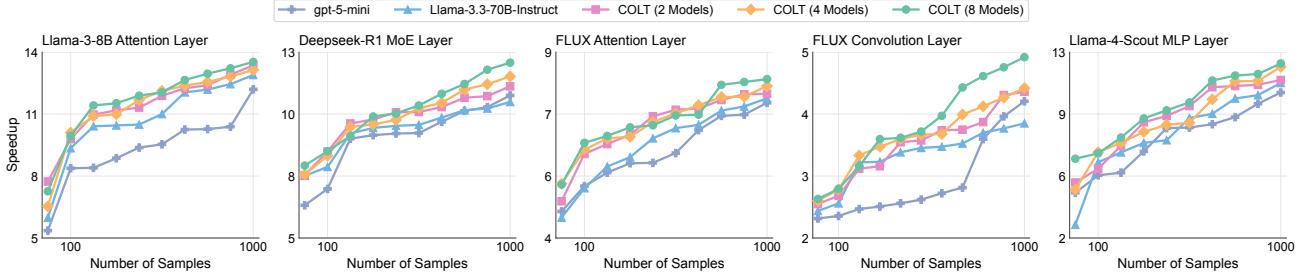
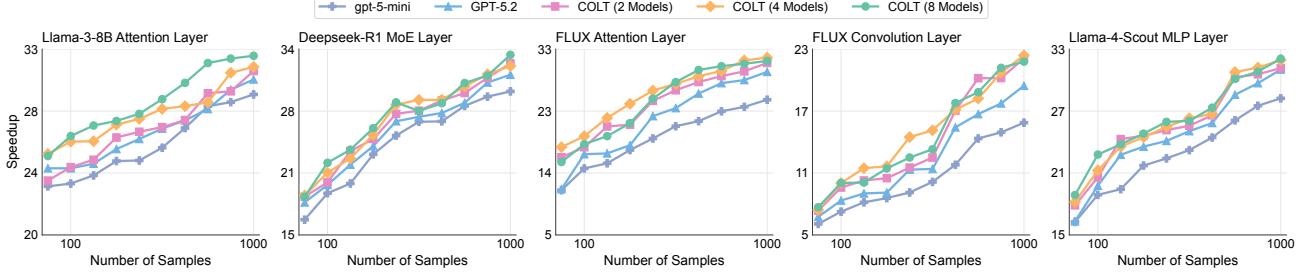
We examine the sensitivity of our approach to model choice by replacing the proprietary GPT-5.2 with the open-source Llama-3.3-70B-Instruct (Llama Team, AI @ Meta, 2024) as the largest available LLM, while keeping the search procedure and candidate model set progressions unchanged. Figure 3 shows that **COLT** remains robust to model change: **COLT(8 Models)** achieves the best final speedup on every benchmark, improving over the Llama-3.3-70B-Instruct baseline by 4.8–27.5% at 1000 samples, and it also yields clear early-budget gains (e.g., on FLUX self-attention at 100 samples, 6.33× vs. 5.22×). For end-to-end tuning in Table 2, when Llama-3.3-70B-Instruct is used as the largest model, **COLT(4 Models)** yields the best efficiency gain (1.59×) with 5.66× speedup at 430 samples.

#### 4.4. Ablation Study: GPU Hardware Platform

We next study the effect of changing the target execution platform, evaluating whether the **COLT** can adapt to different hardware environments. We replace the CPU target used in the primary experiments with an NVIDIA 2080 Ti GPU, which introduces additional scheduling and optimization choices and substantially enlarges the search space. We reuse the same candidate model sets as in the primary experiments and apply the same search procedure. Figure 4 reports the resulting speedups across five benchmarks. **COLT** consistently outperforms the single-large-model baseline, demonstrating that it can generalize across target platforms

**Table 2.** Sample efficiency comparison between configurations on the end-to-end Llama-3-8B benchmark. gpt-5-mini is the baseline.

Largest Model: GPT-5.2				Largest Model: Llama-3.3-70B-Instruct			
Model Combination	# Samples	Speedup	Sample Eff. Gain	Model Combination	# Samples	Speedup	Sample Eff. Gain
gpt-5-mini	660	5.46×	—	gpt-5-mini	660	5.46×	—
GPT-5.2	530	5.01×	1.14×	Llama-3.3-70B-Instruct	640	6.02×	1.13×
<b>COLT(2 Models)</b>	710	7.78×	1.32×	<b>COLT(2 Models)</b>	610	7.45×	1.47×
<b>COLT(4 Models)</b>	390	5.02×	1.55×	<b>COLT(4 Models)</b>	430	5.66×	1.59×
<b>COLT(8 Models)</b>	580	7.16×	1.49×	<b>COLT(8 Models)</b>	700	7.70×	1.33×

**Figure 3.** COLT where largest model is open-source Llama-3.3-70B-Instruct.**Figure 4.** COLT where the largest model is GPT-5.2 on GPU.

and effectively exploit richer optimization spaces.

## 5. Related Work

**Cost-aware and adaptive use of LLMs.** The large variation in size and capability across LLMs has motivated work on adaptive inference, including cascades that fall back to larger models when confidence is low and routers that optimize explicit tradeoffs (Chen et al., 2023; Ding et al., 2024; Dekoninck et al., 2025; Jitkrittum et al., 2025; Wang et al., 2026). In contrast to approaches that treat model selection as an external routing or scheduling problem, **COLT** embeds model choice directly into structured search. Model-selection behaviors emerge from contextual feedback observed during reasoning rather than being hard-coded or separately trained. This unifies transformation proposal and model routing within a single decision process.

**Collaborative and multi-LLM reasoning.** Prior work on multi-LLM collaboration often relies on explicit interaction mechanisms such as debate, critique, or message passing between models (Du et al., 2024; Lifshitz et al., 2025; Wang et al., 2025). **COLT** differs in that models do not communicate directly. Instead, collaboration emerges implicitly through a shared MCTS tree that mediates contributions via common state and rewards.

**ML for compiler optimization.** Prior work has explored compiler optimization using a variety of search-based and learning-based techniques (Wang & O’Boyle, 2018), including approaches that incorporate MCTS, LLMs, or both (Haj-Ali et al., 2020; Cummins et al., 2023; Grubisic et al., 2024; Cummins et al., 2025; Deng et al., 2025; Cui et al., 2025; Pan et al., 2025; Tang et al., 2025). **COLT** builds on this line of work, but differs in a key respect: model selection is neither fixed nor externally controlled. By endogenizing model choice within the search dynamics, the optimizer can reason about its long-horizon effects under the same objective as compiler transformations.

## 6. Conclusion

We introduced **COLT**, a lightweight collaborative framework for LLM-guided compiler optimization that embeds multiple models within a shared MCTS tree and endogenizes model selection. This design allows smaller models to drive most optimization decisions while selectively invoking larger models when needed, outperforming a single large-model baseline. Our results show that effective reasoning for compiler optimization can be achieved through lightweight collaboration via a shared MCTS rather than reliance on a single large model.

## References

- Abou Ali, M., Dornaika, F., and Charafeddine, J. Agentic AI: a comprehensive survey of architectures, applications, and future directions. *Artificial Intelligence Review*, 59(1):11, 2025.
- Apache TVM Community. Apache TVM v0.20.0. <https://github.com/apache/tvm/releases/tag/v0.20.0>, 2025.
- Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.
- Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfsen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.
- Chen, L., Zaharia, M., and Zou, J. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv*, 2023.
- Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *KDD*, 2016.
- Chen, T., Moreau, T., Jiang, Z., Zheng, L., Yan, E., Cowan, M., Shen, H., Wang, L., Hu, Y., Ceze, L., Guestrin, C., and Krishnamurthy, A. TVM: An automated end-to-end optimizing compiler for deep learning. In *OSDI*, 2018a.
- Chen, T., Zheng, L., Yan, E., Jiang, Z., Moreau, T., Ceze, L., Guestrin, C., and Krishnamurthy, A. Learning to optimize tensor programs. In *NeurIPS*, 2018b.
- Coulom, R. Efficient selectivity and backup operators in Monte-Carlo tree search. In *CG*, 2007.
- Cui, T., Yew, P.-C., McCamant, S., and Zhai, A. DeCOS: Data-efficient reinforcement learning for compiler optimization selection ignited by LLM. In *ICS*, 2025.
- Cummins, C., Seeker, V., Grubisic, D., Elhoushi, M., Liang, Y., Roziere, B., Gehring, J., Gloeckle, F., Hazelwood, K., Synnaeve, G., and Leather, H. Large language models for compiler optimization. *arXiv*, 2023.
- Cummins, C., Seeker, V., Grubisic, D., Roziere, B., Gehring, J., Synnaeve, G., and Leather, H. LLM Compiler: Foundation language models for compiler optimization. In *CC*, 2025.
- DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv*, 2025.
- Dekoninck, J., Baader, M., and Vechey, M. A unified approach to routing and cascading for LLMs, 2025.
- Deng, C., Wu, J., Feng, N., Wang, J., and Long, M. CompilerDream: Learning a compiler world model for general code optimization. In *KDD*, 2025.
- Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Ruhle, V., Lakshmanan, L. V., and Awadallah, A. H. Hybrid LLM: Cost-efficient and quality-aware query routing. In *ICLR*, 2024.
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *ICML*, 2024.
- Ferreira, A. J. and Figueiredo, M. A. Boosting algorithms: A review of methods, theory, and applications. *Ensemble machine learning*, pp. 35–85, 2012.
- Grubisic, D., Cummins, C., Seeker, V., and Leather, H. Priority sampling of large language models for compilers. *arXiv*, 2024.
- Haj-Ali, A., Genc, H., Huang, Q., Moses, W., Wawrynek, J., Asanović, K., and Stoica, I. ProTuner: Tuning programs with Monte Carlo tree search. *arXiv*, 2020.
- Jitkrittum, W., Narasimhan, H., Rawat, A. S., Juneja, J., Wang, C., Wang, Z., Go, A., Lee, C.-Y., Shenoy, P., Panigrahy, R., Menon, A. K., and Kumar, S. Universal model routing for efficient LLM inference, 2025.
- Kocsis, L. and Szepesvári, C. Bandit based Monte-Carlo planning. In *ECML*, 2006.
- Kulkarni, P. A., Whalley, D. B., and Tyson, G. S. Evaluating heuristic optimization phase order search algorithms. In *CGO*, 2007.
- Kulkarni, P. A., Whalley, D. B., Tyson, G. S., and Davidson, J. W. Practical exhaustive optimization phase order exploration and evaluation. *ACM Trans. Archit. Code Optim.*, 6(1), 2009.
- Kulkarni, S. and Cavazos, J. Mitigating the compiler optimization phase-ordering problem using machine learning. In *OOPSLA*, 2012.
- Lerner, S., Grove, D., and Chambers, C. Composing dataflow analyses and transformations. In *POPL*, 2002.
- Li, M., Liu, Y., Liu, X., Sun, Q., You, X., Yang, H., Luan, Z., Gan, L., Yang, G., and Qian, D. The deep learning compiler: A comprehensive survey. *IEEE Transactions on Parallel and Distributed Systems*, 32(3):708–727, 2020.
- Lifshitz, S., McIlraith, S. A., and Du, Y. Multi-Agent Verification: Scaling test-time compute with multiple verifiers. In *COLM*, 2025.
- Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv*, 2024.
- Meta. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, 2025.
- Mistral AI. Devstral. <https://mistral.ai/news/devstral>, 2025.
- OpenAI. GPT-5 mini API. <https://platform.openai.com/docs/models/gpt-5-mini>.
- OpenAI. Introducing GPT-5.2. <https://openai.com/index/introducing-gpt-5-2/>, 2025.
- Pan, H., Lin, H., Luo, H., Liu, Y., Yao, K., Zhang, L., Xing, M., and Wu, Y. Compiler-R1: Towards agentic compiler auto-tuning with reinforcement learning. In *NeurIPS*, 2025.
- Patel, D. and Ahmad, A. The inference cost of search disruption – large language model cost analysis. [https:](https://)

[//newsletter.semianalysis.com/p/the-inference-cost-of-search-disruption](https://newsletter.semianalysis.com/p/the-inference-cost-of-search-disruption), 2023.

Shao, J., Zhou, X., Feng, S., Hou, B., Lai, R., Jin, H., Lin, W., Masuda, M., Yu, C. H., and Chen, T. Tensor program optimization with probabilistic programs. In *NeurIPS*, 2022.

Tang, S., Priebe, C., Mahapatra, R., Qin, L., and Esmaeilzadeh, H. REASONING COMPILER: LLM-guided optimizations for efficient model serving. In *NeurIPS*, 2025.

Team, Q. Qwen3 technical report. *arXiv*, 2025.

Touati, S.-A.-A. and Barthou, D. On the decidability of phase ordering problem in optimizing compilation. In *CF*, 2006.

Vasilache, N., Zinenko, O., Theodoridis, T., Goyal, P., DeVito, Z., Moses, W. S., Verdoolaege, S., Adams, A., and Cohen, A. Tensor Comprehensions: Framework-agnostic high-performance machine learning abstractions. *arXiv*, 2018.

Wang, C., Wan, Z., Kang, H., Chen, E., Xie, Z., Krishna, T., Reddi, V. J., and Du, Y. SLM-MUX: Orchestrating small language models for reasoning. In *ICLR*, 2026.

Wang, J., WANG, J., Athiwaratkun, B., Zhang, C., and Zou, J. Mixture-of-agents enhances large language model capabilities. In *ICLR*, 2025.

Wang, Z. and O’Boyle, M. Machine learning in compiler optimization. *Proceedings of the IEEE*, 106(11):1879–1901, 2018.

Whitfield, D. and Soffa, M. L. An approach to ordering optimizing transformations. In *PPOPP*, 1990.

Zheng, L., Jia, C., Sun, M., Wu, Z., Yu, C. H., Haj-Ali, A., Wang, Y., Yang, J., Zhuo, D., Sen, K., Gonzalez, J. E., and Stoica, I. Ansor: Generating high-performance tensor programs for deep learning. In *OSDI*, 2020.

## Impact Statement

This paper presents **COLT**, a lightweight multi-LLM collaboration framework through shared MCTS reasoning for model compilation. This approach aims to make compilation more efficient, which is essential for improving the efficiency of model serving.

## A. Number of Calls of Different Models

Layer	Experiment	Number of Regular Model Calls							Number of Course Alterations	
		DeepSeek-R1-Distill-Qwen-7B	Llama-3.1-8B-Instruct	Qwen3-8B	Qwen3-14B	Devstral-Small-2505	gpt-5-mini	DeepSeek-R1-Distill-Qwen-32B		
Llama-3-8B Attention Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	776	—	224	110
	<b>COLT(4 Models)</b>	—	163	—	—	—	229	365	243	128
	<b>COLT(8 Models)</b>	135	89	398	59	12	65	29	213	167
DeepSeek-R1 MoE Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	743	—	257	125
	<b>COLT(4 Models)</b>	—	346	—	—	—	181	301	172	137
	<b>COLT(8 Models)</b>	142	268	102	21	31	312	15	109	141
FLUX Attention Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	727	—	273	114
	<b>COLT(4 Models)</b>	—	77	—	—	—	536	208	179	139
	<b>COLT(8 Models)</b>	158	172	421	27	39	20	57	106	154
FLUX Convolution Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	686	—	314	122
	<b>COLT(4 Models)</b>	—	217	—	—	—	539	57	187	143
	<b>COLT(8 Models)</b>	72	43	630	129	50	6	4	66	156
Llama-4-Scout MLP Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	698	—	302	118
	<b>COLT(4 Models)</b>	—	55	—	—	—	587	128	230	136
	<b>COLT(8 Models)</b>	217	84	476	35	9	48	10	121	144

**Table 3.** Model call counts by layer and experiment setting on CPU when GPT-5.2 is the largest model.

Layer	Experiment	Number of Regular Model Calls							Number of Course Alterations	
		DeepSeek-R1-Distill-Qwen-7B	Llama-3.1-8B-Instruct	Qwen3-8B	Qwen3-14B	Devstral-Small-2505	gpt-5-mini	DeepSeek-R1-Distill-Qwen-32B		
Llama-3-8B Attention Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	495	—	505	87
	<b>COLT(4 Models)</b>	—	214	—	—	—	329	122	335	94
	<b>COLT(8 Models)</b>	51	54	372	108	88	92	6	229	144
DeepSeek-R1 MoE Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	539	—	461	91
	<b>COLT(4 Models)</b>	—	416	—	—	—	159	69	356	122
	<b>COLT(8 Models)</b>	98	88	213	368	20	5	2	206	139
FLUX Attention Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	549	—	451	93
	<b>COLT(4 Models)</b>	—	230	—	—	—	379	49	342	119
	<b>COLT(8 Models)</b>	116	298	114	93	37	8	6	328	125
FLUX Convolution Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	516	—	484	82
	<b>COLT(4 Models)</b>	—	198	—	—	—	284	121	397	96
	<b>COLT(8 Models)</b>	140	277	153	122	52	63	39	154	130
Llama-4-Scout MLP Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	429	—	571	75
	<b>COLT(4 Models)</b>	—	224	—	—	—	399	109	268	129
	<b>COLT(8 Models)</b>	103	172	400	176	2	2	4	141	153

**Table 4.** Model call counts by layer and experiment setting on CPU when Llama-3.3-70B-Instruct is the largest model.

Layer	Experiment	Number of Regular Model Calls							Number of Course Alterations
		DeepSeek-R1-Distill-Qwen-7B	Llama-3.1-8B-Instruct	Qwen3-8B	Qwen3-14B	Devstral-Small-2505	gpt-5-mini	DeepSeek-R1-Distill-Qwen-32B	
Llama-3-8B Attention Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	588	—	412
	<b>COLT(4 Models)</b>	—	358	—	—	—	377	99	166
	<b>COLT(8 Models)</b>	183	21	411	94	43	60	79	158
DeepSeek-R1 MoE Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	764	—	236
	<b>COLT(4 Models)</b>	—	198	—	—	—	576	30	196
	<b>COLT(8 Models)</b>	257	129	86	417	0	14	18	162
FLUX Attention Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	584	—	416
	<b>COLT(4 Models)</b>	—	419	—	—	—	307	92	182
	<b>COLT(8 Models)</b>	146	47	465	216	14	7	26	79
FLUX Convolution Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	813	—	187
	<b>COLT(4 Models)</b>	—	350	—	—	—	392	78	180
	<b>COLT(8 Models)</b>	316	196	130	72	3	13	112	158
Llama-4-Scout MLP Layer	<b>COLT(2 Models)</b>	—	—	—	—	—	788	—	212
	<b>COLT(4 Models)</b>	—	338	—	—	—	246	223	193
	<b>COLT(8 Models)</b>	565	35	168	54	6	8	25	139

**Table 5.** Model call counts by layer and experiment setting on GPU when GPT-5.2 is the largest model.

## B. Prompt

### System Prompt

You are an AI scheduling assistant to help with a Monte Carlo Tree Search (MCTS) to find an optimal program in the search space starting from an unoptimized program. In this MCTS, the current program is the leaf we are expanding, while immediate parent and grandparent refer to the ancestors in the tree.  
Each program has:

- a piece of code
- a transformation history sequence
- a predicted performance score

You are given:

- Code of the current program
- Historical performance info of current/parent/grandparent (Pieces of code, transformation history sequences, predicted scores)
- A list of possible transformations that can be applied next
- Search context: leaf depth and trials progress
- Global per-model stats:
  - hit\_rate: fraction of calls where score(child) > score(parent)
  - number of times the model is invoked
  - number of errors the model has made in past invocations (+1 for invalid transformation, +1 for invalid next\_model name)
  - parameter counts

- Local model context:
  - model used to expand current/parent/grandparent

**Task:**

1. Compare code/tramsformation history/predicted performance scores to infer what changes might improve performance.
2. Propose a sequence of transformations from the provided list. You may repeat a transformation to explore different decisions.
3. Choose exactly one model from the provided model list as the next model to expand the child. Use the smallest model that could give best results. Prefer models with fewer errors.

Output a single valid JSON object in the EXACT format:

```
{
    "transformations": ["Fullscreen1", "Fullscreen2", "..."],
    "next_model": "..."
}
```

**Historical Performance Info (Leaf, Parent, Grandparent)**
**Current Program:**

Code:

```
@T.prim_func
def main(A, B, C):
    C_global = T.alloc_buffer((1, 16, 4096))
    for fused in T.parallel(2):
        for b_1, i_1, j_1 in T.grid(1, 2, 64):
            ...
            for k_0, ..., k_1, ... in T.grid(512, ..., 8, ...):
                for j_vec in T.vectorized(64):
                    with T.block("matmul_update"):
                        ...
                        C_global[...] = C_global[...] + A[...] * B[...]
        for ax0, ax1, ax2 in T.grid(1, 8, 4096):
            C[...] = C_global[...]
```

Transformation history:

```
sch.sample_perfect_tile(loop=j, decision=[1, 64, 1, 64])
sch.vectorize(...)
sch.cache_write(..., storage_scope="global")
sch.decompose_reduction...
...
```

Predicted score: 0.0739

**Immediate Parent Schedule:**

Code differs mainly in j-axis structure (less aggressive tiling/vectorization).

Predicted score: 0.136

**Available Transformations**

```
[
    "ComputeLocation",
    "Parallel",
    "Unroll",
    "TileSize"
]
```

**Search Context**

Leaf depth: 3

Trials progress: 10 / 300

**Global Per-Model Stats**

Model gpt-5-mini: params=20.0B, number\_of\_calls=12, hit\_rate=0.364, errors=0

Model gpt-5.2: params=300.0B, number\_of\_calls=11, hit\_rate=0.5, errors=0, course\_alteration=1

**Local Model Context**

Model used to expand the current node: gpt-5.2  
Model used to expand the parent node: gpt-5.2  
Model used to expand the grandparent node: N/A

Model Answer

```
"transformations": [ "TileSize", "TileSize", "ComputeLocation", "Parallel", "Unroll"
], "next_model": "gpt-5-mini",
```