

Reasoning as Gradient: Scaling MLE Agents Beyond Tree Search

Yifei Zhang^{1*†}, Xu Yang^{2*}, Xiao Yang^{2*},
Bowen Xian², Qizheng Li^{3†}, Shikai Fang⁴, Jingyuan Li^{5†},
Jian Wang², Mingrui Xu^{6†}, Yuge Zhang², Weiqing Liu^{2‡}, Jiang Bian²

¹Nanjing University, ²Microsoft Research Asia, ³Peking University,
⁴Zhejiang University, ⁵Wuhan University, ⁶Hong Kong University of Science and Technology
 [GitHub Repo](#)  [GPT-5 Execution Traces](#)

Abstract

LLM-based agents for machine learning engineering (MLE) predominantly rely on tree search, a form of gradient-free optimization that uses scalar validation scores to rank candidates. As LLM reasoning capabilities improve, exhaustive enumeration becomes increasingly inefficient compared to directed updates, analogous to how accurate gradients enable efficient descent over random search. We introduce GOME, an MLE agent that operationalizes gradient-based optimization. GOME maps structured diagnostic reasoning to gradient computation, success memory to momentum, and multi-trace execution to distributed optimization. Under a closed-world protocol that isolates architectural effects from external knowledge, GOME achieves a state-of-the-art 35.1% any-medal rate on MLE-Bench with a restricted 12-hour budget on a single V100 GPU. Scaling experiments across 10 models reveal a critical crossover: with weaker models, tree search retains advantages by compensating for unreliable reasoning through exhaustive exploration; as reasoning capability strengthens, gradient-based optimization progressively outperforms, with the gap widening at frontier-tier models. Given the rapid advancement of reasoning-oriented LLMs, this positions gradient-based optimization as an increasingly favorable paradigm. We release our codebase and GPT-5 traces at: <https://github.com/microsoft/RD-Agent>.

1 Introduction

Automating Machine Learning Engineering (MLE) remains a long-standing challenge. Given a dataset and an evaluation metric, the objective is to autonomously complete the end-to-end development

pipeline, including data preprocessing, feature engineering, model selection, and hyperparameter tuning, to deliver a high-performing predictive system. With rapid advances in code generation and complex reasoning capabilities of large language models (Jaech et al., 2024; Guo et al., 2025), LLM-based MLE agents have recently emerged (Jiang et al., 2025; Liu et al., 2025; Nam et al., 2025), demonstrating the ability to interpret task descriptions, diagnose data characteristics, implement complete solutions, and iteratively refine them using execution feedback. To enable systematic comparison of these agents, MLE-Bench (Chan et al., 2024) provides a standardized benchmark of 75 Kaggle competitions.

Current MLE agents predominantly adopt search-based exploration. AIDE (Jiang et al., 2025) first introduced this design motivated by two practical considerations: constrained context windows that made chain-based prompting prone to prompt bloat, and multi-branch exploration to mitigate uncertainty in single-shot generations. Subsequent work largely follows this paradigm: ML-Master (Liu et al., 2025) incorporates Monte Carlo Tree Search, and AIRA (Toledo et al., 2025) proposes graph-based exploration. Importantly, these methods use *directed* topologies (e.g., UCT/PUCT-style node selection with value backpropagation). Our characterization of “gradient-free” is therefore about the *optimization signal*, not the search topology: improvement is still primarily selected by scalar rewards/scores rather than by explicitly reasoning about update direction from rich execution feedback. Under this signal-level view, prior search-based agents exhibit two limitations. First, they are driven by **score-centric selection**: rich execution feedback (error traces, training dynamics, detailed logs) is compressed into scalar rewards that mainly decide which node/branch survives, discarding diagnostic information needed to determine how to update. This information loss becomes increasingly

*Equal contribution.

†Work done during an internship at Microsoft Research Asia.

‡Correspondence: Weiqing.Liu@microsoft.com

costly as LLM reasoning capabilities improve, since stronger models could otherwise extract precise improvement directions from such feedback. Second, these methods operate over a **predefined action space**: the agent selects among fixed templates, which cannot capture the effectively continuous nature of code modifications and may mismatch the specific failure modes revealed by execution.

We argue that MLE inherently favors gradient-based optimization. Unlike domains punctuated by hard dead ends, MLE pipelines are typically repairable: improvements come from iteratively refining the same pipeline rather than discarding intermediate states. Moreover, the space of code modifications is effectively continuous, ranging from minor hyperparameter tweaks to architectural changes, favoring methods that propose state-conditioned updates based on execution feedback, rather than searching over a fixed action vocabulary. The key question is when such updates become preferable to enumeration. This trade-off is well understood in classical optimization: gradient-based methods dominate when gradient estimates are accurate, but degrade when noisy (Maheswaranathan et al., 2019; Asadi et al., 2024). In the MLE agent setting, the LLM’s reasoning capability determines gradient signal quality. When reasoning is weak, multi-branch enumeration hedges against unreliable proposals; as reasoning-oriented models advance,¹ the balance should shift toward gradient-based optimization: **gradient-based methods should increasingly outperform tree search as reasoning capability improves** (Figure 1).

To test this claim, we propose **GOME** (Gradient-based Optimization for Machine Learning Engineering), an agent that adopts gradient-based optimization in place of enumeration-driven search. GOME uses a chain-based iterative framework where each step updates the current solution along LLM-generated improvement directions, organized into modular components that mirror classical optimizer modules (Table 2). We evaluate under a **closed-world protocol** to isolate architectural effects from knowledge augmentation: agents use only task-provided materials and execution feedback, without external knowledge retrieval. Under this protocol, GOME achieves a **35.1%** any-medal rate on MLE-BENCH with a **12-hour** budget on a single **V100** GPU, surpassing

¹From o1-preview (Jaech et al., 2024) to DeepSeek-R1 (Guo et al., 2025), o3 (OpenAI, 2025), and GPT-5 (Singh et al., 2025).

the previous state-of-the-art held by search-based methods. Component ablations confirm that each optimizer-inspired module contributes meaningfully. Scaling analysis, which compares GOME against tree-search baselines (including an MCTS variant sharing GOME’s codebase and previous state-of-the-art search-based methods) across 10 models ranging from non-reasoning (GPT-4o) to frontier reasoning models (GPT-5), further validates our claim: while tree-search methods show advantages with weaker reasoners, GOME’s advantage grows with model capability. Given the rapid advancement of reasoning-oriented LLMs (Sun et al., 2025; Xu et al., 2025), this positions gradient-based optimization as an increasingly favorable paradigm.

In summary, this work makes the following key contributions: (1) We propose GOME, an MLE agent that adopts gradient-based optimization rather than score-centric candidate ranking, establishing a functional mapping between agent components and classical optimizer modules. (2) GOME achieves state-of-the-art performance on MLE-BENCH under closed-world evaluation. Scaling analysis across 10 models validates our claim: GOME’s advantage over tree search grows with model reasoning capability, confirming that stronger reasoning enables more effective gradient-based optimization. (3) We release our codebase and GPT-5 execution traces to support reproducibility and contribute to the community.

2 Related Work

LLM-based MLE Agents. Current MLE agents explore various search topologies: tree search (AIDE (Jiang et al., 2025), ML-Master (Liu et al., 2025), KompeteAI (Kulibaba et al., 2025)), graph-based frameworks (AIRA (Toledo et al., 2025), InternAgent-MLE (Du et al.)), and evolutionary methods (FM Agent (Li et al., 2025)). Some approaches augment search with external knowledge retrieval from Kaggle notebooks, arXiv papers, or expert knowledge bases. MLE-STAR (Nam et al., 2025) adopts a chain structure with web retrieval and targeted code-block refinement; however, its iteration remains score-driven, using ablation-based scoring to select which block to refine rather than reasoning about how to update (see Appendix D.3). We summarize key design choices in Table 1. Notably, AIRA formalizes research agents as $(\mathcal{F}, \pi_{\text{sel}}, \mathcal{O}, \pi_{\text{op}}, \tau)$ and shows that, under the AIDE operator set, changing the search policy alone yields only limited gains, suggesting that

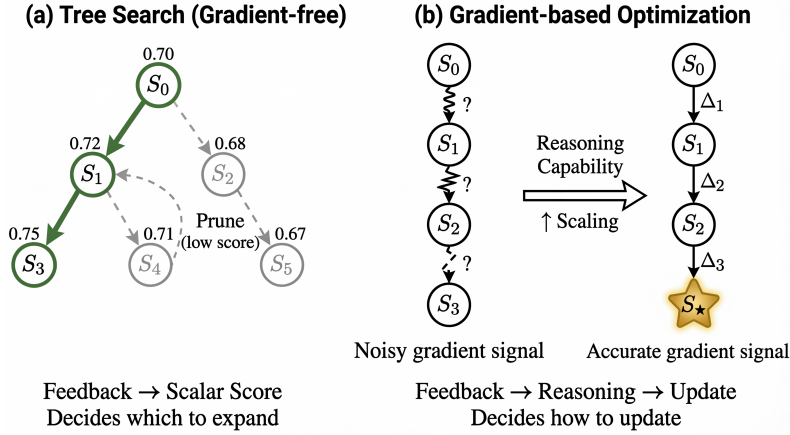


Figure 1: Two paradigms for utilizing execution feedback. (a) Score-centric: feedback is compressed into scalar rewards that guide node/branch selection (which may itself be directed, e.g., via UCT/PUCT). (b) Gradient-based: structured reasoning over execution feedback determines how to update the current solution; stronger reasoning yields more accurate signals.

operator design can be a major bottleneck beyond search topology (Toledo et al., 2025). InternAgent-MLE (Du et al.) further extends tree search to graph search with cross-branch trajectory reuse, yet the optimization signal remains score-driven. These examples illustrate a common pattern: topologies grow increasingly directed, but feedback utilization stays score-centric. In contrast, GOME integrates scalar scores with execution feedback for update: structured reasoning over these signals generates improvement hypotheses that directly guide solution modification.

Method	Structure	Feedback Role	Ext. Know.	Code
AIDE	Tree	Ranking	✗	✓
ML-Master	Tree	Ranking	✗	✓
AIRA	Graph	Ranking	✗	✓
MLE-STAR	Chain	Ranking	✓	✓
InternAgent-MLE	Graph	Ranking	✓	✗
KompeteAI	Tree	Ranking	✓	✗
FM Agent	Evolutionary	Ranking	✓	✗
GOME	Chain	Update	✗	✓

Table 1: Comparison of LLM-based MLE agents. Feedback Role indicates how execution feedback drives the primary optimization loop: Ranking denotes score-centric selection among candidates; Update denotes diagnostic reasoning over structured feedback to generate state-conditioned code modifications.

Reasoning as Optimization. A growing line of work operationalizes LLM reasoning as an optimization signal, using task outcomes to iteratively refine artifacts rather than exhaustively enumerating candidates (Yang et al., 2023; Shinn et al., 2023; Zhang et al., 2025). In prompt optimization, LLMs propose edits conditioned on performance

feedback, with methods ranging from discrete optimization procedures (Pryzant et al., 2023; Cui et al., 2024) to population-based evolution (Guo et al., 2023). In agentic settings, trial outcomes are converted into self-reflections that guide behavior updates (Madaan et al., 2023), while textual “gradients” formalize refinement as structured edits (Yuksekgonul et al., 2024; Cheng et al., 2024). While this paradigm has been widely adopted in prompt tuning and general-purpose agent tasks, it remains underexplored in MLE, where existing methods still rely on score-based search (Table 1). GOME bridges this gap, instantiating reasoning-as-optimization for MLE with structured feedback as the gradient signal for code updates.

3 GOME

We formulate MLE tasks as finding an optimal solution within the code space under resource constraints.

$$s^* = \arg \max_{s \in \mathcal{S}} h(\mathcal{T}, s) \quad \text{s.t.} \quad \text{cost}(s) \leq B, \quad (1)$$

where \mathcal{S} is the space of valid ML pipelines, $h(\mathcal{T}, s)$ evaluates solution s on task \mathcal{T} (e.g., accuracy), and B bounds the total computational budget.

Existing MLE agents typically organize \mathcal{S} as a tree or graph and use scalar scores to guide expansion. We propose GOME, which instead focuses on how to update. Inspired by gradient-based optimization, GOME extracts directional improvement signals from structured feedback via LLM reasoning. While true gradients are undefined in discrete code space, LLM reasoning serves a functionally

analogous role: analyzing execution results to determine not just whether a solution improved, but why it improved and what to change next. Table 2 summarizes this analogy.

Figure 2 illustrates the framework. GOME initializes N parallel optimization traces, each beginning with a distinct hypothesis (i.e., an improvement direction for the ML pipeline) through forced diversification (§3.5). Each trace maintains local best solution $s^{*(i)}$ and experiment history $\mathcal{H}^{(i)}$, while synchronizing via a shared success memory \mathcal{M} . Each iteration proceeds through four stages: (1) **Execution**: run current solution and collect feedback (§3.1); (2) **Validation**: apply hierarchical checks to form structured feedback $f_t^{(i)}$ (§3.2); (3) **Memory update**: contribute successful hypotheses to \mathcal{M} (Momentum accumulation) (§3.3); (4) **Reasoning**: generate the next hypothesis $\eta_{t+1}^{(i)}$ by combining local feedback with shared memory (Gradient computation) (§3.4). This local-global interplay enables collaborative optimization (§3.5), with full algorithmic details in Appendix C.

3.1 Execution Feedback

Each iteration begins with a trace executing its current solution and collecting local feedback. Executing solution $s_t^{(i)}$ on trace i produces execution feedback:

$$(\text{perf}_t, \text{trace}_t) = \text{Execute}(s_t^{(i)}, s^{*(i)}), \quad (2)$$

where $s^{*(i)}$ denotes the best solution found by trace i so far. The component $\text{perf}_t = (h_t, h^{*(i)}) \in \mathbb{R}^2$ contains current score and trace-local best score as scalar metrics; trace_t captures execution logs (stdout/stderr, runtime logs) and code diffs relative to the trace’s current best solution $s^{*(i)}$, enabling diagnosis of implementation issues.

3.2 Hierarchical Validation

A critical challenge in LLM-based optimization is distinguishing genuine improvements from deceptive ones. Solutions may achieve high validation scores through shortcuts such as data leakage or metric gaming. Based on execution feedback, hierarchical validation determines acceptance decision $d_t^{(i)} \in \{0, 1\}$ and diagnostic reason_t :

$$(d_t^{(i)}, \text{reason}_t) = \text{Validate}(s_t^{(i)}, \text{perf}_t, \text{trace}_t), \quad (3)$$

forming the structured feedback $f_t^{(i)} = (\text{perf}_t, \text{trace}_t, \text{reason}_t)$. The validation process

enforces three sequential stages: (1) **Format correctness**: verifies output schema via rule-based checks; (2) **Evaluation alignment**: detects data leakage and overfitting risks via LLM analysis, e.g., accessing test labels or future information; (3) **Comprehensive analysis**: verifies whether the hypothesis achieved its intended effect and assesses code quality. The LLM makes the final accept decision $d_t^{(i)}$ based on this reasoning when $h(s_t^{(i)}) \gtrsim h(s^{*(i)})$, with reason_t capturing the analysis from whichever stage determined the outcome, informing subsequent hypothesis generation. Empirically, hierarchical validation detects 66.7% of deceptive overfitting attempts, compared to 0% for purely score-driven baselines (Appendix D.1).

3.3 Success Memory

Upon validation, the hypothesis and its feedback are committed to the shared success memory \mathcal{M} :

$$\mathcal{M}_{t+1} = \begin{cases} \mathcal{M}_t \cup \{(\eta_t^{(i)}, f_t^{(i)}, \Delta h_t^{(i)})\} & \text{if } d_t^{(i)} = 1 \\ \mathcal{M}_t & \text{otherwise} \end{cases} \quad (4)$$

where $\eta_t^{(i)}$ is the hypothesis that led to solution $s_t^{(i)}$, $f_t^{(i)}$ is the structured feedback (§3.2), and $\Delta h_t^{(i)} = h_t^{(i)} - h^{*(i)}$ records the score difference (may be negative if accepted for code quality merits). Each memory entry contains the complete context: what was tried (hypothesis), what happened (feedback), and how much it helped (score delta). \mathcal{M} is distinct from the raw local history $\mathcal{H}^{(i)}$: in single-trace settings, it serves as a curated subset of verified successes; in multi-trace settings, it naturally extends to a global repository aggregating discoveries across all workers. Functionally analogous to momentum, this global memory biases future updates toward proven directions.

3.4 Structured Reasoning

The reasoning module generates the next improvement hypothesis, serving as the gradient signal that directs optimization. It first extracts challenges $\mathcal{C}_t^{(i)}$ from structured feedback $f_t^{(i)}$ and local history $\mathcal{H}^{(i)}$, shifting from exploratory analysis (early iterations) to targeted diagnosis (addressing specific execution failures). For each identified challenge c , the module generates a concrete hypothesis η_c by conditioning on the trace’s complete local state $(s^{*(i)}, f_t^{(i)}, \mathcal{H}^{(i)})$.

Concept	GOME Component	Functional Role
Gradient ∇L	Structured Reasoning	Signal: decides how to update
Momentum	Success Memory	Acceleration via proven patterns
Distributed SGD	Multi-trace Optimization	Parallelism with knowledge sharing

Table 2: Analogy between GOME and gradient-based optimization.

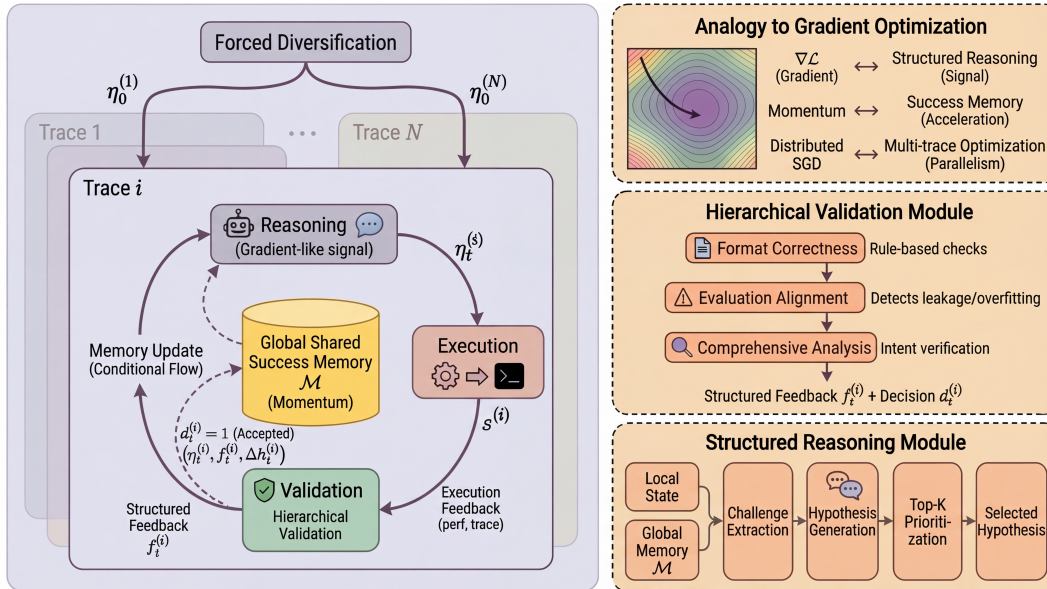


Figure 2: Overview of the GOME framework. Multiple traces optimize in parallel, synchronizing through a global shared success memory \mathcal{M} . Each trace iteratively executes solutions, validates improvements, updates shared memory, and reasons over local-global feedback to generate the next hypothesis. Right panels illustrate the gradient-based optimization analogy and key module details.

Candidates are scored across dimensions \mathcal{D} (impact, alignment, novelty, feasibility, risk-reward)²:

$$\text{score}(\eta) = \sum_{d \in \mathcal{D}} w_d \cdot \text{score}_d(\eta, \mathcal{M}), \quad (5)$$

where the success memory \mathcal{M} (§3.3) modulates scoring: hypotheses resembling past successes gain confidence, while similar failures receive penalties. Rather than greedily selecting the top-scored hypothesis, we sample from top- k to maintain exploration diversity (in multi-trace settings, cross-trace mechanisms further guide selection, see §3.5). The selected hypothesis is finally implemented as solution $s_{t+1}^{(i)}$ via code generation and iterative refinement.

3.5 Multi-trace Optimization

Single-trace optimization risks local optima and cannot recover from poor early decisions. GOME employs N parallel traces that synchronize via \mathcal{M} , enabling online knowledge sharing where trace i 's

successful discoveries immediately inform traces $j \neq i$.

Forced Diversification at Initialization. To prevent redundant exploration (which renders memory sharing ineffective), we maximize initial search coverage by strictly enforcing diversity. Trace n 's starting hypothesis is conditioned on all prior proposals to ensure orthogonality:

$$\eta_0^{(n)} = \text{GenerateHypothesis}(\mathcal{T}, \{\eta_0^{(j)}\}_{j < n}), \quad (6)$$

forcing the swarm to investigate distinct regions of the solution space from the outset.

Cross-trace Hypothesis Selection. In multi-trace settings, each trace constructs a candidate pool from three complementary sources:

$$\mathcal{H}_{\text{cand}} = \{\eta_c\}_{c \in \mathcal{C}_t^{(i)}} \cup \{\eta^*\} \cup \text{Sample}(\mathcal{M}), \quad (7)$$

where $\{\eta_c\}$ are hypotheses generated by the current trace (§3.4), η^* is the hypothesis from the highest-scoring iteration in \mathcal{M} , and $\text{Sample}(\mathcal{M})$ retrieves

²Weights: (0.4, 0.2, 0.2, 0.1, 0.1) respectively.

hypotheses via a probabilistic kernel considering embedding similarity and score diffs. An LLM-based selector then produces the final hypothesis:

$$\eta_{t+1}^{(i)} = \text{LLMSelect}(\mathcal{H}_{\text{cand}}, \{f_\eta\}_{\eta \in \mathcal{M}}), \quad (8)$$

where $\{f_\eta\}$ denotes the associated feedback stored alongside each hypothesis. The selector has flexibility to select, modify, or generate based on the candidate pool and historical outcomes. Additional details of the probabilistic interaction kernel are provided in Appendix C.3.

3.6 Robust Implementation

We employ several mechanisms to ensure reliable optimization: (1) **Development-evaluation separation**: initial iterations run on a small data subset with rapid debug loops, where failed code triggers iterative debugging that analyzes error messages and proposes targeted fixes; final validation executes on full data. (2) **Multi-seed selection**: for final submission, we rerun top- k candidates with multiple seeds and select the best-performing result, reducing variance from stochastic LLM outputs. (3) **Adaptive time management**: an LLM-based module extends time budget based on task complexity and failure patterns, avoiding premature termination on difficult tasks.

4 Experiments

We evaluate GOME on MLE-Bench under a rigorous closed-world protocol to validate our claim that gradient-based optimization scales more effectively than search as reasoning capability improves. Additional experiments including protocol comparison and cost analysis are provided in Appendix A. Following prior MLE-Bench evaluations, we adopt any-medal rate as the primary metric. We additionally report improvement rate and validation-test correlation (IC) to characterize per-step optimization dynamics.

4.1 Experiment Setup

Benchmark. All experiments are conducted on MLE-Bench (Chan et al., 2024), a comprehensive benchmark comprising 75 Kaggle competitions for evaluating AI agents on machine learning engineering tasks, categorized by complexity into low, medium, and high. MLE-Bench-Lite consists of 22 low-complexity tasks for efficient evaluation. We report any-medal rate as the primary indicator.

Implementation. We evaluate GOME using three frontier LLMs representing a spectrum of reasoning capabilities: DeepSeek-R1 (Guo et al., 2025), o3 (OpenAI, 2025), and GPT-5 (Singh et al., 2025), ordered by increasing reasoning strength. For the test environment, we use 12 vCPUs, 220GB RAM, and 1 NVIDIA V100 GPU with a 12-hour time budget. All results are averaged over three runs with different random seeds.

Baselines. We benchmark against leading closed-world MLE agents on MLE-Bench, including MLAB (Huang et al., 2023), OpenHands (Wang et al., 2024), AIDE (Jiang et al., 2025), AIRA (Toledo et al., 2025), and ML-Master (Liu et al., 2025). For ML-Master, we re-evaluate under identical hardware and time budget to ensure fairness; for other baselines, we use results reported in their respective papers or the official leaderboard.

4.2 Main Results

GOME establishes a new state-of-the-art on MLE-Bench. As shown in Table 3, GOME outperforms all closed-world baselines, achieving a peak any-medal rate of **35.1%** with GPT-5. Beyond aggregate performance, GOME excels in solution quality: it achieves a 96.0% valid submission rate and converts 16.4% of tasks to Gold medals. Under identical constraints (12h, V100), GOME matches ML-Master on weaker reasoners (DeepSeek-R1: 23.4% vs. 22.7%) but pulls ahead significantly with stronger models, widening the gap to **+11.1 percentage points** with GPT-5 and improving Gold rate. This widening gap is consistent with our hypothesis (validated across 10 models in Section 5). Furthermore, GOME matches the performance of AIRA despite operating under stricter resource constraints (**half the time**, weaker GPU), highlighting the efficiency of gradient-based optimization (full cross-protocol comparisons are provided in Table 8).

Accurate gradient signals drive rapid convergence. On MLE-Bench-Lite (Table 4), GOME achieves a **68.2%** medal rate, matching the SOTA open-world method (Leeroo) despite lacking external retrieval. This confirms that for tractable tasks, internal diagnostics provide high-fidelity gradients sufficient for rapid convergence (Nesterov et al., 2018). Conversely, the plateau on high-complexity tasks across all methods indicates a “noisy gradient” regime where reasoning limits are reached, suggesting that further gains require restoring signal accuracy through stronger models or external

Agent	Time	GPU	Medal rate in different complexity (%)				Other evaluation dimensions (%)		
			Low	Medium	High	All	Valid	Median+	Gold
MLAB									
gpt-4o-24-08	24h	A10 [†]	4.2±1.5	0.0±0.0	0.0±0.0	1.3±0.5	44.3±2.6	1.9±0.7	0.8±0.5
OpenHands									
gpt-4o-24-08	24h	A10 [†]	11.5±3.4	2.2±1.3	1.9±1.9	5.1±1.3	52.0±3.3	7.1±1.7	2.7±1.1
AIDE									
gpt-4o-24-08	24h	A10 [†]	19.0±1.3	3.2±0.5	5.6±1.0	8.6±0.5	54.9±1.0	14.4±0.7	5.0±0.4
o1-preview	24h	A10 [†]	34.3±2.4	8.8±1.1	10.0±1.9	16.9±1.1	82.8±1.1	29.4±1.3	9.4±0.8
AIRA									
o3	24h	H200	55.0±1.5	22.0±1.2	<u>21.7±1.1</u>	31.6±0.8	<u>97.5±0.3</u>	45.5±0.8	17.3±0.4
ML-Master									
DeepSeek-R1	12h	V100	47.0±6.6	9.7±2.3	20.0±3.8	22.7±0.8	91.6±0.4	32.9±0.4	12.4±0.4
o3	12h	V100	42.4±4.0	12.3±1.8	20.0±0.0	22.7±1.3	94.2±0.4	31.5±2.2	13.3±0.8
GPT-5	12h	V100	51.5±3.0	10.5±2.6	17.8±2.2	24.0±2.3	98.2±0.4	34.7±0.4	12.4±1.6
GOME (ours)									
DeepSeek-R1	12h	V100	50.0±0.0	9.7±0.9	20.0±0.0	23.4±0.4	88.9±0.4	35.1±0.4	12.4±0.9
o3	12h	V100	<u>59.1±4.5</u>	20.2±0.9	22.2±2.2	<u>32.5±1.2</u>	94.2±0.4	43.1±0.9	17.3±0.8
GPT-5	12h	V100	68.2±2.6	<u>21.1±1.5</u>	22.2±2.4	35.1±0.4	96.0±0.0	<u>45.3±0.0</u>	<u>16.4±0.8</u>

Table 3: Percentage of achieving any medals across different ML task complexity levels (left) and other evaluation dimensions (right) on MLE-Bench. Reporting results are mean ± SEM over 3 seeds. Valid, Median+, and Gold indicate the percentage of submissions with valid score, above median score, and gold medal. Best performances are marked in **bold**; second best are underlined. [†]GPU configuration from official MLE-Bench evaluation setup.

Agent	Any-Medal (%)
KompeteAI [†] (Gemini-2.5-flash,6h)	51.5±1.5
AIRA (o3,24h)	55.0±1.5
InternAgent-MLE [†] (DeepSeek-R1,12h)	62.1±3.0
FM Agent [†] (Gemini-2.5-pro,24h)	62.1±1.5
Thesis [†] (GPT-5-Codex,24h)	65.2±1.5
Leeroo [†] (Gemini-3-pro,24h)	68.2±2.6
GOME (GPT-5,12h)	68.2±2.6

Table 4: Performance comparison on MLE-Bench-Lite. Best performances are marked in **bold**. [†]Methods using open-world protocol with external resource access.

knowledge.

Beyond benchmark evaluation, we validate GOME on a live Kaggle competition, demonstrating its ability to discover non-trivial feature engineering strategies autonomously (Appendix D.4).

4.3 Ablation Study

To validate GOME’s design, we ablate its three core components: (1) **w/o Structured Reasoning**, replacing diagnostic analysis with simple error-correction prompts; (2) **w/o Success Memory**, removing the global repository of proven patterns; and (3) **w/o Multi-trace Optimization**, reducing to single-trajectory optimization without cross-trace knowledge sharing.

As shown in Table 5, removing any component

compromises the system’s overall robustness, particularly the primary Any-Medal rate. The **Structured Reasoning** ablation causes the most severe degradation, with improvement rate dropping from 41.1% to 22.6%, indicating that without diagnostic analysis, most optimization steps fail to produce valid improvements. Removing **Success Memory** shows a different failure mode: while per-iteration success remains decent, the lack of momentum from proven patterns leads to redundant exploration, reducing medal rate by 6.2%. Finally, **Multi-trace Optimization** is essential for diversity: its removal degrades final performance despite high local improvement rates, suggesting that cross-trace sharing helps the swarm escape the local optima that trap single-trace agents. All three components contribute meaningfully (Table 2).

5 Scaling Analysis

To validate our claim that gradient signal quality depends on reasoning capability, we evaluate GOME across three tiers: Efficiency (e.g., GPT-4o-mini), Advanced (e.g., DeepSeek-R1), and Frontier (e.g., o3, GPT-5)³. We compare against GOME-MCTS, a controlled variant replacing gradient updates with tree search within identical prompt and infrastruc-

³See Appendix A.1 for complete tier definitions and per-model results.

Configuration	Medal	Gold	Impr.	IC
GOME (full)	35.1	16.4	41.1	0.92
w/o Structured Reasoning	25.8	13.3	22.6	0.83
w/o Success Memory	28.9	16.9	36.2	0.87
w/o Multi-trace Optimization	32.4	15.1	41.3	0.88

Table 5: Ablation study on GOME components using GPT-5 under 12-hour budget (averaged over 3 runs). Impr.: improvement rate, the percentage of iterations where the proposed solution passes hierarchical validation. IC: Spearman correlation between validation and test improvements.

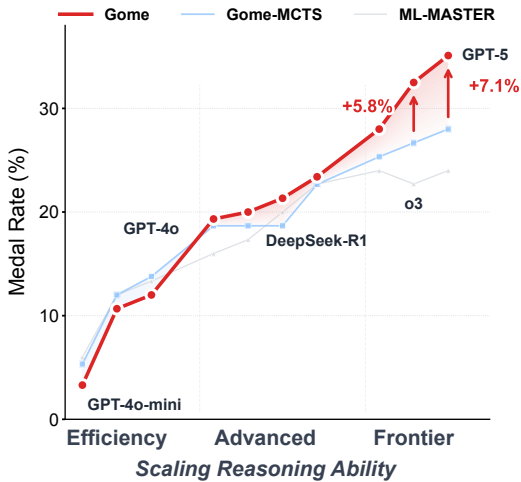


Figure 3: GOME performance scales with model reasoning capability. As base model capability increases from Efficiency to Frontier tiers, GOME’s advantage over search-based baselines widens significantly.

ture, isolating the impact of the optimization strategy (implementation details in Appendix B.3).

Figure 3 reveals a distinct phase transition. On Efficiency-tier models, GOME lags behind: when gradient signals are noisy due to weak reasoning, exhaustive search outperforms directed updates. However, a **crossover** emerges as capability increases—search gains plateau while GOME accelerates, widening the gap to +5.8% on o3 and +7.1% on GPT-5. Figure 4 traces this divergence temporally. GOME exhibits rapid early convergence across all tiers, but sustainability differs: on weaker models (GPT-4o), unreliable gradients cause premature convergence, allowing MCTS to eventually catch up; on Frontier models, accurate gradient signals enable GOME to maintain its lead throughout, achieving both faster convergence and superior final performance. Complete per-model results and tier-wise gap visualization are provided in Appendix A.1.

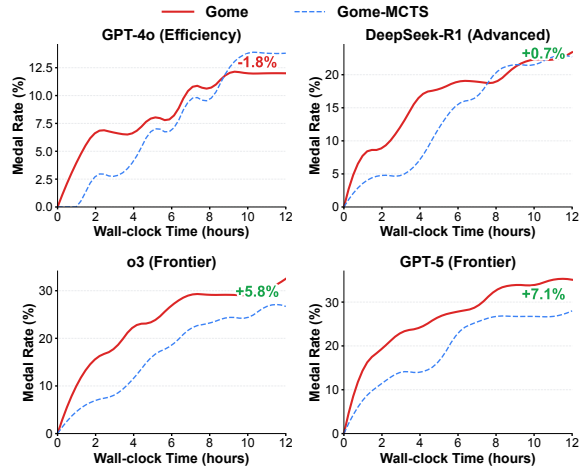


Figure 4: Optimization dynamics across model capability tiers. GOME (solid red) exhibits rapid early convergence, while MCTS (dashed blue) starts slower but catches up on weaker models. On Frontier-tier models, GOME maintains its advantage throughout.

These results validate our central claim: gradient-based optimization increasingly outperforms tree search as reasoning capability improves. More broadly, they reveal a fundamental divergence in scaling properties: tree search scales with *inference compute* (traversing more nodes), while gradient-based optimization scales with *model capability* (reasoning better). Given the rapid advancement of reasoning-oriented LLMs, this positions gradient-based optimization as the increasingly favorable paradigm for MLE agents.

6 Conclusion

We presented GOME, reframing MLE agents through gradient-based optimization. By treating LLM reasoning as gradient signals, success memory as momentum, and multi-trace execution as distributed optimization, GOME establishes a principled correspondence between classical optimization and agentic ML engineering.

Our experiments demonstrate state-of-the-art performance on MLE-Bench, while our scaling analysis reveals a more fundamental finding: tree search scales with inference compute (traversing more nodes), while GOME scales with model capability through increasingly accurate gradient signals. As foundation models advance in reasoning ability, this positions gradient-based optimization as an increasingly favorable paradigm.

This work opens a new design axis for MLE agents: rather than engineering more sophisticated search strategies, future systems may benefit from

investing in gradient quality through richer feedback signals and stronger diagnostic reasoning. Future work may also explore hybrid strategies that combine tree search for early-stage exploration with gradient-based refinement as reasoning quality improves within a run. We release our codebase and GPT-5 execution traces to support reproducibility and future research.

Limitations

Dependency on Reasoning Capability. GOME’s effectiveness is sensitive to the base model’s reasoning capability. As shown in our scaling analysis, GOME underperforms tree-search methods on efficiency-tier models, where weaker reasoners may hallucinate diagnostic feedback. This also implies higher inference costs when using frontier models (Table 9). However, this dependency aligns with the trajectory of LLM development: as reasoning-oriented models advance rapidly, the capability threshold for effective gradient-based optimization will become increasingly accessible.

Evaluation Scope. We evaluated under a closed-world protocol on MLE-Bench, restricting agents to task-provided materials without external knowledge retrieval. While this ensures rigorous comparison of optimization strategies by controlling for confounding factors such as retrieval quality and knowledge source coverage, it limits generalizability: real-world MLE tasks often benefit from external knowledge, and GOME’s performance in open-world settings remains less thoroughly validated. Additionally, real Kaggle competitions allow participants to access shared notebooks and public baselines, so our closed-world setting may underestimate the performance achievable in practical scenarios.

Lack of Formal Convergence Guarantees. Our gradient-based framing establishes functional correspondence rather than mathematical equivalence. Unlike true gradients in continuous optimization, LLM-generated update signals lack formal properties such as convergence guarantees. This means that in rugged landscapes where solutions require fundamentally different paradigms rather than incremental edits, gradient-based guidance may get trapped in local optima. In practice, this is mitigated by the nature of MLE tasks: ML pipelines are typically repairable, making the smoothness assumption reasonable for most practical scenarios.

Our multi-trace optimization and success memory further mitigate this risk by enabling shared exploration across trajectories.

References

- Artificial Analysis Team. 2025. *Artificial Analysis: Independent Benchmarks and Performance Landscape of AI Models*. Accessed: 2025-12-28.
- Saeed Asadi, Sonia Gharibzadeh, Hajar Kazemi Naeini, Masoud Reihanifar, Morteza Rahimi, Shiva Zangeneh, Aseel Smerat, and Lazim Abdullah. 2024. Comparative analysis of gradient-based optimization techniques using multidimensional surface 3d visualizations and initial point sensitivity. *arXiv preprint arXiv:2409.04470*.
- Léon Bottou, Frank E Curtis, and Jorge Nocedal. 2018. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311.
- Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, and 1 others. 2024. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*.
- Ching-An Cheng, Allen Nie, and Adith Swaminathan. 2024. Trace is the next autodiff: Generative optimization with rich feedback, execution traces, and llms. *Advances in Neural Information Processing Systems*, 37:71596–71642.
- Anthony Cui, Pranav Nandyalam, Andrew Rufail, Ethan Cheung, Aiden Lei, Kevin Zhu, and Sean O’Brien. 2024. Introducing mapo: Momentum-aided gradient descent prompt optimization. *arXiv preprint arXiv:2410.19499*.
- Shangheng Du, Xiangchao Yan, Dengyang Jiang, Jiakang Yuan, Yusong Hu, Xin Li, Liang He, Bo Zhang, and LEI BAI. Internagent-mle: Navigating fine-grained optimization for coding agent.
- Timur Garipov, Pavel Izmailov, Dmitrii Podoprikin, Dmitry P Vetrov, and Andrew G Wilson. 2018. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31.
- Paul Gauthier. 2025. Aider LLM Leaderboards: Code Editing and Refactoring Benchmarks. <https://aider.chat/docs/leaderboards/>. Accessed: 2025-12-28.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

- Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Evoprompt: Connecting llms with evolutionary algorithms yields powerful prompt optimizers. *arXiv e-prints*, pages arXiv-2309.
- Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*.
- Akira Isihara. 2013. *Statistical physics*. Academic Press.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. 2025. Aide: Ai-driven exploration in the space of code. *arXiv preprint arXiv:2502.13138*.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.
- Stepan Kulibaba, Artem Dzhalilov, Roman Pakhomov, Oleg Svidchenko, Alexander Gasnikov, and Aleksei Shpilman. 2025. Kompetelai: Accelerated autonomous multi-agent system for end-to-end pipeline generation for machine learning problems. *arXiv preprint arXiv:2508.10177*.
- Annan Li, Chufan Wu, Zengle Ge, Yee Hin Chong, Zhinan Hou, Lizhe Cao, Cheng Ju, Jianmin Wu, Huaiming Li, Haobo Zhang, and 1 others. 2025. The fm agent. *arXiv preprint arXiv:2510.26144*.
- Zexi Liu, Yuzhu Cai, Xinyu Zhu, Yujie Zheng, Runkun Chen, Ying Wen, Yanfeng Wang, Siheng Chen, and 1 others. 2025. Ml-master: Towards ai-for-ai via integration of exploration and reasoning. *arXiv preprint arXiv:2506.16499*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.
- Niru Maheswaranathan, Luke Metz, George Tucker, Dami Choi, and Jascha Sohl-Dickstein. 2019. Guided evolutionary strategies: Augmenting random search with surrogate gradients. In *International Conference on Machine Learning*, pages 4264–4273. PMLR.
- Jaehyun Nam, Jinsung Yoon, Jiefeng Chen, Jinwoo Shin, Sercan Ö Arık, and Tomas Pfister. 2025. Mle-star: Machine learning engineering agent via search and targeted refinement. *arXiv preprint arXiv:2506.15692*.
- Yurii Nesterov and 1 others. 2018. *Lectures on convex optimization*, volume 137. Springer.
- Fan Nie, Junlin Wang, Harper Hua, Federico Bianchi, Yongchan Kwon, Zhenting Qi, Owen Queen, Shang Zhu, and James Zou. 2026. Dsgym: A holistic framework for evaluating and training data science agents. *arXiv preprint arXiv:2601.16344*.
- OpenAI. 2025. *Introducing openai o3 and o4-mini*. Accessed: 2025-12-22.
- Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.
- Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, and 1 others. 2025. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*, 57(11):1–43.
- Edan Toledo, Karen Hambardzumyan, Martin Josifoski, Rishi Hazra, Nicolas Baldwin, Alexis Audran-Reiss, Michael Kuchnik, Despoina Magka, Minqi Jiang, Alisia Maria Lupidi, and 1 others. 2025. Ai research agents for machine learning: Search, exploration, and generalization in mle-bench. *arXiv preprint arXiv:2507.02554*.
- Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangu Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, and 1 others. 2024. Openhands: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*.
- Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.
- Mert Yuksekogonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Wenlin Zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, and 1 others. 2025. Process vs. outcome reward: Which is better for agentic rag reinforcement learning. *arXiv preprint arXiv:2505.14069*.

Appendix

A More Experimental Analysis

A.1 Extended Scaling Analysis

This section provides detailed per-model results supporting the scaling analysis in Section 5. We categorize models into three capability tiers following the stratification established by the Artificial Analysis Intelligence Index⁴, SWE-bench, and Aider leaderboard⁵ (Artificial Analysis Team, 2025; Jimenez et al., 2023; Gauthier, 2025): **Efficiency** models are optimized for throughput and cost; **Advanced** models represent balanced capability-efficiency trade-offs; **Frontier** models define current state-of-the-art reasoning.

Model	GOME	GOME-MCTS	ML-Master	Runs
Efficiency Tier Models				
GPT-4o-mini	3.3	5.3	6.0	2
GPT-5-nano	10.7	12.0	12.0	2
GPT-4o	12.0	13.8	13.3	3
Advanced Tier Models				
GPT-5-mini	19.3	18.7	16.0	2
Grok-4	20.0	18.7	17.3	2
Grok-4.1	21.3	18.7	20.0	2
DeepSeek-R1	23.4	22.7	22.7	3
Frontier Tier Models				
DeepSeek-v3.2	28.0	25.3	24.0	2
o3	32.5	26.7	22.7	3
GPT-5	35.1	28.0	24.0	3

Table 6: Complete per-model scaling results. All values are any-medal rates (%) on MLE-Bench. **Bold** indicates the best performance per model. Runs indicates the number of independent trials.

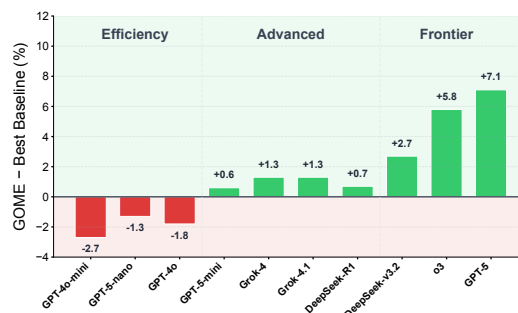


Figure 5: Performance gap between GOME and the best baseline across model capability tiers. Negative values (red) indicate GOME underperforms; positive values (green) indicate GOME outperforms. The crossover occurs at the Advanced tier, with the gap widening progressively in the Frontier tier.

Table 6 and Figure 5 report the complete results for all ten models evaluated in our scaling

⁴<https://artificialanalysis.ai/>

⁵<https://aider.chat/docs/leaderboards/>

analysis. On Efficiency-tier models, GOME underperforms the best baseline consistently, with gaps ranging from -1.3% to -2.7% . This confirms that weaker models produce unreliable gradient signals. The high sensitivity of gradient-based optimization causes it to be confidently misguided by hallucinated diagnostics, while search-based methods remain robust through exhaustive exploration. The crossover emerges in the Advanced tier, where all four models show positive gaps ranging from $+0.6\%$ to $+1.3\%$. The advantage is modest but consistent, suggesting that mid-capability models begin producing sufficiently accurate gradient signals to offset the exploration benefits of tree search. At the Frontier tier, GOME’s advantage becomes substantial and the gap widens progressively from $+2.7\%$ to $+7.1\%$, demonstrating that stronger reasoners amplify GOME’s relative gains rather than merely maintaining them.

These results validate our central claim: gradient-based optimization exhibits fundamentally different scaling properties from tree search. The performance gap transitions from negative to positive and continues to widen as model capability increases, implying that gradient-based optimization will become increasingly favorable as foundation models advance.

A.2 Additional Ablations

Our main experiments use a 12-hour time budget (half of MLE-Bench’s standard 24-hour allocation) for computational efficiency. To examine whether GOME continues to improve with the full time budget, we select o3 and GPT-5 as representative Frontier-tier models and extend to 24 hours.

Model	12h	24h	Improvement
o3	32.5	36.4	+3.9 (+12.0%)
GPT-5	35.1	40.4	+5.3 (+15.1%)

Table 7: GOME performance under half-time (12h) and full-time (24h) budgets (averaged over 3 runs). All values are any-medal rates (%) on MLE-Bench. See Table 8 for full results.

Table 7 shows that GOME continues to benefit from additional optimization time without plateauing. Both models achieve substantial gains when doubling the time budget, with o3 improving by 12.0% and GPT-5 by 15.1% relatively. Notably, the stronger model (GPT-5) exhibits larger improvements, consistent with our scaling claim: higher-

quality gradient signals enable more sustained optimization progress. This suggests that gradient-based optimization has not yet saturated its potential on current Frontier models, and further gains may be achievable with increased compute budgets.

A.3 Impact of External Knowledge Access

Table 3 presents results under two evaluation protocols. Methods marked with \checkmark operate in a closed-world setting, where agents cannot access external knowledge sources during evaluation but may employ carefully designed prompts, multi-stage pipelines, and framework-level optimizations. Methods marked with \times operate in an open-world setting, where agents can additionally leverage external resources such as web search, documentation retrieval, and curated solution libraries. We include both for comprehensive comparison.

Rationale for Closed-World Evaluation. Our main experiments (Section 4) adopt the closed-world protocol for three reasons. First, it isolates architectural contributions from confounding factors, ensuring that performance differences can be attributed to the optimization framework rather than retrieval quality or knowledge source coverage. Second, it enables fair comparison across methods, as open-world results depend heavily on which external APIs are available, how retrieved content is filtered, and the quality of curated knowledge bases. Third, it directly tests our core claim that improved LLM reasoning translates to better gradient signals, whereas external knowledge access would obscure this relationship by providing an alternative path to performance gains.

Performance Analysis. External knowledge access shows varying impact across task complexity. On low-complexity tasks, GOME achieves the highest medal rate across both protocols, suggesting that strong reasoning can fully compensate for the absence of external resources on well-defined problems. The gap widens on medium and high-complexity tasks, where open-world methods benefit from domain-specific knowledge retrieval for specialized competitions. Despite this inherent disadvantage, GOME remains highly competitive, achieving top-three overall performance while operating under strictly closed-world conditions with shorter time budgets and weaker hardware than leading open-world methods.

Agent	w/o Ext. Know.	Time	GPU	Medal rate in different complexity (%)			
				Low	Medium	High	All
MLAB							
gpt-4o-24-08	✓	24h	A10 [‡]	4.2±1.5	0.0±0.0	0.0±0.0	1.3±0.5
OpenHands							
gpt-4o-24-08	✓	24h	A10 [‡]	11.5±3.4	2.2±1.3	1.9±1.9	5.1±1.3
AIDE							
gpt-4o-24-08	✓	24h	A10 [‡]	19.0±1.3	3.2±0.5	5.6±1.0	8.6±0.5
o1-preview	✓	24h	A10 [‡]	34.3±2.4	8.8±1.1	10.0±1.9	16.9±1.1
AIRA							
o3	✓	24h	H200	55.0±1.5	22.0±1.2	21.7±1.1	31.6±0.8
ML-Master							
DeepSeek-R1	✓	12h	V100	47.0±6.6	9.7±2.3	20.0±3.8	22.7±0.8
o3	✓	12h	V100	42.4±4.0	12.3±1.8	20.0±0.0	22.7±1.3
GPT-5	✓	12h	V100	51.5±3.0	10.5±2.6	17.8±2.2	24.0±2.3
KompeteAI							
Gemini-2.5-flash	✗	6h	A100	51.5±1.5	–	–	–
Neo							
Multi Frontier-Tier Models	✗	36h	A100	48.5±1.5	29.8±2.3*	24.4±2.2	34.2±0.9
InternAgent-MLE							
DeepSeek-R1	✗	12h	A800	62.1±3.0	26.3±2.6	24.4±2.2	36.4±1.2
MLE-STAR							
Gemini-2.5-pro	✗	12h	8*V100	63.6±6.0	–	–	–
FM Agent							
Gemini-2.5-pro	✗	12h	A800	62.1±1.5	<u>36.8±1.5</u>	33.3±0.0	<u>43.6±0.8</u>
Thesis							
GPT-5-Codex	✗	24h	H100	65.2±1.5*	45.6±7.2	<u>31.1±2.2</u>	48.4±3.6
GOME (ours)							
DeepSeek-R1	✓	12h	V100	50.0±0.0	9.7±0.9	20.0±0.0	23.4±0.4
o3	✓	12h	V100	59.1±4.5	20.2±0.9	22.2±2.2	32.5±1.2
o3	✓	24h	V100	63.6±2.6	27.2±2.3	24.4±2.2	36.4±0.4
GPT-5	✓	12h	V100	<u>68.2±2.6</u>	21.1±1.5	22.2±2.4	35.1±0.4
GPT-5	✓	24h	V100	71.2±1.5	28.1±0.9	26.7±0.0*	40.4±0.9*

Table 8: Percentage of achieving any medals across different ML task complexity levels on MLE-Bench. Reporting results are mean ± SEM over 3 seeds. Best performances are marked in **bold**; second best are underlined; third best are marked with *. “–” indicates results not reported in the original work. [‡]GPU configuration from official MLE-Bench evaluation setup.

A.4 Cost Analysis

Table 9 reports API pricing and average cost per competition for Frontier-tier models under the 12-hour evaluation budget.

Model	Input	Output	Cost/Task	Medal Rate
DeepSeek-v3.2	\$0.28 (\$0.028)	\$0.42	\$1.89	28.0%
GPT-5	\$1.25	\$10.00	\$20.74	35.1%
o3	\$2.00	\$8.00	\$26.53	32.5%

Table 9: API pricing (per million tokens) and average cost per competition for Frontier-tier models. Cost/Task is averaged over 3 independent runs. Parenthetical value indicates cache read price.

DeepSeek-v3.2 achieves 28.0% medal rate at \$1.89 per task, offering the best cost-efficiency among Frontier models. Its pricing is an order of magnitude lower than alternatives, with cache read further reducing input cost to \$0.028/M tokens for repeated context. GPT-5 achieves the highest performance (35.1%) at \$20.74 per task, where the high output price (\$10/M tokens) dominates the cost. o3 incurs the highest total cost (\$26.53) while underperforming GPT-5, likely due to extended reasoning traces that increase token consumption without proportional performance gains. These results suggest that model selection should balance performance requirements against budget constraints, with DeepSeek-v3.2 as a strong choice for cost-sensitive deployments.

A.5 Task-type Breakdown

Table 10 presents detailed score comparison across 9 representative MLE-Bench tasks, grouped by data modality. We evaluate GOME with three frontier-tier backbone models: DeepSeek-R1 (Guo et al., 2025), o3 (OpenAI, 2025), and GPT-5 (Singh et al., 2025).

While GPT-5 achieves significantly higher overall medal rate (35.1%) compared to o3 (32.5%) and DeepSeek-R1 (23.4%), the task-level scores reveal a more nuanced picture. GPT-5 achieves the best score on 4 out of 9 tasks, but o3 also wins 4 tasks, and DeepSeek-R1 achieves the top score on plant-pathology despite having the lowest overall medal rate. This discrepancy arises because medal rate measures the proportion of tasks exceeding competition-specific thresholds, not absolute score superiority. GPT-5’s advantage lies in consistently crossing medal thresholds across more tasks, rather than dominating every individual competition. The results suggest that different backbone models bring complementary strengths, and GOME’s framework

effectively harnesses these diverse capabilities regardless of the underlying LLM.

B Implementation Details

B.1 GOME Hyperparameters

Table 11 summarizes the hyperparameter configuration used in GOME experiments. Parameters are organized according to the framework components described in Section 3.

B.2 ML-Master Reproduction

We reproduce ML-Master (Liu et al., 2025) as a representative state-of-the-art tree search baseline. We obtain the official implementation from the authors’ public repository⁶ and adapt it to our evaluation protocol.

Environment Configuration. The original ML-Master experiments were conducted on NVIDIA A100 GPUs. We adapt the implementation to run on NVIDIA V100 GPUs within our standardized Docker environment, ensuring fair comparison across all methods. All experiments use the same MLE-Bench Docker image used for GOME, eliminating potential confounds from environment differences. We apply a 12-hour wall-clock time budget per task, consistent with our closed-world evaluation protocol.

Model Substitution. The original ML-Master implementation uses DeepSeek-R1 and GPT-4o as backbone models. To enable scaling analysis across model capability tiers, we modify the inference module to support configurable LLM backends while preserving all other components (tree construction, node selection, and expansion strategies). This allows us to evaluate ML-Master with models ranging from GPT-4o-mini to GPT-5.

Hyperparameters. We adopt the default hyperparameters from the original implementation, including search depth, branching factor, and selection criteria. No task-specific tuning is performed to ensure a fair comparison with GOME, which also uses fixed hyperparameters across all tasks.

Reproduction Validation. Table 12 compares our reproduced results with those reported in the original paper. Minor performance differences are attributable to hardware differences (V100 vs. A100), Docker environment variations, stochastic

⁶<https://github.com/sjtu-sai-agents/ML-Master>

Task	Metric	DeepSeek-R1	o3	GPT-5
Image Tasks				
aptos2019-blindness-detection	QWK \uparrow	0.8938	0.9243	0.9158
inaturalist-2019-fgvc6	Top-1 Error \downarrow	0.2951	0.3201	0.2261
plant-pathology-2021-fgvc8	F1 \uparrow	0.9100	0.8949	0.9044
Text Tasks				
detecting-insults-in-social-commentary	AUC \uparrow	0.9384	0.9452	0.9176
jigsaw-toxic-comment-classification	AUC \uparrow	0.9848	0.9856	0.9865
spooky-author-identification	Logloss \downarrow	0.3291	0.2880	0.2293
Tabular Tasks				
nomad2018-predict-transparent-conductors	RMSLE \downarrow	0.0615	0.0593	0.0618
stanford-covid-vaccine	MCRMSE \downarrow	0.2289	0.2263	0.3365
tabular-playground-series-dec-2021	Accuracy \uparrow	0.9615	0.9631	0.9632

Table 10: Score comparison on 9 MLE-Bench tasks across three data modalities. Best result for each task is highlighted in **bold**. \uparrow : higher is better; \downarrow : lower is better.

variation across runs, and our stricter closed-world protocol that prohibits external knowledge retrieval.

Model	Hardware	Medal Rate	Source
DeepSeek-R1	A100	29.3	Official
DeepSeek-R1	V100	22.7	Our Rerun

Table 12: ML-Master reproduction validation on DeepSeek-R1. Original results are from (Liu et al., 2025) on A100 GPUs; ours are averaged over 3 independent runs under closed-world protocol on V100 GPUs. All values are any-medal rates (%) on MLE-Bench with 12-hour budget.

B.3 GOME-MCTS Implementation

Our modular framework naturally supports diverse exploration strategies beyond gradient-based optimization. To provide a fair comparison between optimization paradigms within the same framework, we implement Monte Carlo Tree Search (MCTS), a widely-adopted algorithm in existing MLE agents such as ML-Master (Liu et al., 2025). This implementation, called GOME-MCTS, serves two purposes: (1) it validates that our framework is a general-purpose platform capable of accommodating different search algorithms, and (2) it isolates the effect of the optimization strategy from other architectural differences when comparing against external baselines.

Architectural Mapping. GOME-MCTS reuses GOME’s execution feedback (§3.1), hierarchical validation (§3.2), and robust implementation modules (§3.6). The key difference lies in how optimization decisions are made: GOME-MCTS replaces structured reasoning (§3.4) with PUCT-based node selection, and replaces success memory (§3.3) with

Q -value estimates accumulated through tree search. This substitution isolates the comparison to the core optimization mechanism while controlling for other architectural factors.

PUCT Selection. We adopt the PUCT (Predictor + UCT) variant with uniform prior $P(n, a) = 1$. The selection score balances exploitation and exploration:

$$U(n, a) = Q(n, a) + c_{\text{puct}} \cdot \frac{\sqrt{N(n)}}{1 + N(n, a)}, \quad (9)$$

where $Q(n, a)$ represents the average reward of taking action a from node n , $N(n, a)$ counts how many times action a has been selected from node n , and $N(n)$ is the total visit count for node n . The parameter $c_{\text{puct}} \in [0, \infty)$ controls the exploration-exploitation trade-off: $c_{\text{puct}} = 0$ yields pure exploitation (greedy selection), while larger values increasingly favor exploration of less-visited actions. We set $c_{\text{puct}} = 1.0$ in all experiments.

Reward Design. We experiment with two reward formulations. The binary reward R_{bin} simply assigns 1 for validated submissions and 0 otherwise:

$$R_{\text{bin}} = \begin{cases} 1, & \text{if validated,} \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

When a feedback score v is available, score-based reward R_{score} provides richer signals:

$$R_{\text{score}} = \begin{cases} +\tanh(v), & \text{if validated and higher is better,} \\ -\tanh(v), & \text{if validated and lower is better,} \\ -1, & \text{if rejected.} \end{cases} \quad (11)$$

Hyperparameter	Description	Default
LLM Configuration		
chat_model	Base LLM for reasoning	GPT-5
temperature	LLM decoding temperature	1.0
max_retry	Maximum API retry attempts	12000
retry_wait_seconds	Wait time between retries (s)	5
full_runtime	Total Running timeout (h)	12
Success Memory (§3.3)		
enable_global_memory	Enable shared success memory \mathcal{M} across traces	True
memory_save_type	Content saved to memory (Full: hypothesis + feedback + score)	Full
Structured Reasoning (§3.4)		
llm_select_hypothesis	Use LLM-based hypothesis selector	True
simple_hypothesis	Enable simplified hypothesis format after initialization	True
unique_hypothesis	Enforce unique hypothesis generation within each trace	True
enable_cross_trace_sharing	Allow hypothesis sampling from other traces via \mathcal{M}	True
Multi-trace Optimization (§3.5)		
max_trace_num	Number of parallel traces N	4
merge_hours	Hours between trace synchronization	3
debugging_semaphore	Max concurrent debugging operations	3
running_semaphore	Max concurrent running operations	3
feedback_semaphore	Max concurrent feedback operations	1
cross_trace_diversity	Enforce diversity across traces at initialization	True
Robust Implementation (§3.6)		
coder_timeout_multiplier	Upper bound for coder timeout scaling	4×
runner_timeout_multiplier	Upper bound for runner timeout scaling	4×
timeout_increase_stage	Initial timeout increase stage	1
timeout_stage_patience	Patience before timeout escalation	2
llm_decide_longer_runtime	Use LLM to decide whether to extend execution time	True
fix_seed_and_split	Fix random seed and data split for reproducibility	True
enable_multi-seed_selection	Rerun top candidates with multiple seeds for final submission	True

Table 11: GOME Hyperparameter Configuration.

The tanh transformation bounds rewards in $[-1, 1]$ while preserving score ordering. We use R_{score} in all experiments.

Backpropagation. After each rollout, rewards are backpropagated to update Q values along the trajectory using standard averaging:

$$Q(n, a) \leftarrow \frac{N(n, a) \cdot Q(n, a) + R}{N(n, a) + 1}. \quad (12)$$

Tree Expansion. Each node expansion generates $k = 3$ candidate solutions using the same LLM backbone as GOME. Unlike GOME’s sequential refinement guided by structured reasoning, GOME-MCTS explores multiple branches in parallel and selects based on accumulated Q -values. The maximum tree depth is set to 10, with early termination upon achieving a medal-qualifying score.

C Algorithm Details

C.1 Algorithm for Execution Feedback and Hierarchical Validation

The detailed procedure for generating execution feedback and performing hierarchical validation is

presented in Algorithm 1. This process ensures that the solution s_t is rigorously evaluated against the trace-local best solution s^* through a multi-stage gate mechanism before being accepted.

C.2 Algorithm for Structured Reasoning

The structured reasoning process, which guides the generation of the next hypothesis η_{t+1} , is outlined in Algorithm 2. This algorithm utilizes an adaptive weighting mechanism and leverages the global success memory \mathcal{M} to prioritize hypotheses based on historical success and scenario analysis.

C.3 Multi-trace Optimization Details

This section provides implementation details for the multi-trace optimization described in §3.5.

C.3.1 Candidate Hypothesis Construction

As described in §3.5, each trace constructs a candidate pool from three complementary sources:

$$\mathcal{H}_{\text{cand}} = \{\eta_c\}_{c \in \mathcal{C}_t^{(i)}} \cup \{\eta^*\} \cup \mathcal{H}_{\text{sim}}, \quad (13)$$

where:

Algorithm 1 Execution Feedback and Hierarchical Validation

```
1: Input: current solution  $s_t$ , trace-local best solution  $s^*$ , task specification  $\mathcal{T}$ 
2: Output: structured feedback  $f_t$ , accept decision  $d_t$ 
3:
4:                                     ▷ Stage 1: Execution Feedback (§3.1)
5:  $h_t \leftarrow \text{Evaluate}(s_t, \mathcal{T})$                                      ▷ current validation score
6:  $\text{perf}_t \leftarrow (h_t, h^*)$                                          ▷ scalar metrics: current and best scores
7:  $\text{trace}_t \leftarrow (\text{CodeDiff}(s_t, s^*), \text{Logs}(s_t))$                  ▷ code diffs and execution logs
8:
9:                                     ▷ Stage 2: Hierarchical Validation (§3.2)
10:                                    ▷ Gate 1: Format correctness (rule-based)
11: if  $\neg \text{FormatValid}(s_t, \mathcal{T})$  then
12:    $\text{reason}_t \leftarrow \text{DiagnoseFormat}(s_t, \mathcal{T})$ 
13:   return  $((\text{perf}_t, \text{trace}_t, \text{reason}_t), 0)$ 
14: end if
15:                                     ▷ Gate 2: Evaluation alignment (LLM-based)
16:  $\text{leakage} \leftarrow \text{DetectLeakage}(s_t, \mathcal{T})$                          ▷ test label access, future info, etc.
17: if  $\text{leakage} \neq \emptyset$  then
18:    $\text{reason}_t \leftarrow \text{DiagnoseAlignment}(s_t, \text{leakage})$ 
19:   return  $((\text{perf}_t, \text{trace}_t, \text{reason}_t), 0)$ 
20: end if
21:                                     ▷ Gate 3: Comprehensive analysis (when  $h_t \gtrsim h^*$ )
22:  $\text{reason}_t.\text{hyp\_verify} \leftarrow \text{VerifyHypothesis}(s_t, s^*, \text{trace}_t)$    ▷ intended effect achieved?
23: if  $h_t \gtrsim h^*$  then
24:    $\text{reason}_t.\text{code\_quality} \leftarrow \text{AnalyzeCode}(s_t, s^*)$            ▷ best practices, efficiency
25: end if
26:
27:                                     ▷ Stage 3: Accept Decision (LLM-based)
28:  $f_t \leftarrow (\text{perf}_t, \text{trace}_t, \text{reason}_t)$ 
29:  $d_t \leftarrow \text{LLMJudge}(f_t)$                                          ▷ accept based on full structured feedback
30: return  $(f_t, d_t)$ 
```

Algorithm 2 Structured Reasoning

```
1: Input: feedback  $f_t$ , local history  $\mathcal{H}$ , global success memory  $\mathcal{M}$ , best-so-far  $s^*$ , task  $\mathcal{T}$ 
2: Output: selected hypothesis  $\eta_{t+1}$ , updated solution  $s_{t+1}$ 
3:
4: ▷ Stage 1: Adaptive Weighting
5:  $n_{\text{succ}} \leftarrow |\{e \in \mathcal{H} : d_e = 1\}|$ 
6:  $n_{\text{fail}} \leftarrow |\mathcal{H}| - n_{\text{succ}}$ 
7:  $\lambda \leftarrow \max(0, 3 - \lfloor (3n_{\text{succ}} + 2n_{\text{fail}})/8 \rfloor)$  ▷ decays with experience
8:
9: ▷ Stage 2: Challenge Extraction
10:  $\mathcal{C}_t \leftarrow \emptyset$ 
11: if  $\lambda > 0$  then ▷ early: scenario-driven
12:    $\mathcal{C}_t \leftarrow \mathcal{C}_t \cup \text{AnalyzeScenario}(s^*, \mathcal{T})$  ▷ SOTA alignment, gap identification
13: end if
14: if  $\lambda < 3$  then ▷ later: feedback-driven
15:    $\mathcal{C}_t \leftarrow \mathcal{C}_t \cup \text{AnalyzeHistory}(f_t, \mathcal{H})$  ▷ explicit issues, persistent errors
16: end if
17:
18: ▷ Stage 3: Hypothesis Generation
19: for  $c \in \mathcal{C}_t$  do
20:    $\eta_c \leftarrow \text{Hypothesize}(c, s^*, f_t, \mathcal{T})$  ▷ target component + reasoning
21: end for
22:
23: ▷ Stage 4: Prioritization
24:  $\mathcal{D} \leftarrow \{\text{impact, alignment, novelty, feasibility, risk}\}$ 
25:  $\mathbf{w} \leftarrow (0.4, 0.2, 0.2, 0.1, 0.1)$ 
26: for  $(c, \eta) \in \mathcal{C}_t$  do
27:    $\text{score}(\eta) \leftarrow \sum_{d \in \mathcal{D}} w_d \cdot \text{score}_d(\eta, \mathcal{M})$ 
28: end for
29:
30: ▷ Stage 5: Hypothesis Selection
31: if single-trace then
32:    $\mathcal{H}_{\text{cand}} \leftarrow \text{TopK}(\{(c, \eta_c)\}, k)$ 
33:    $\eta_{t+1} \leftarrow \text{Sample}(\mathcal{H}_{\text{cand}})$  ▷ stochastic selection
34: else ▷ multi-trace (§3.5)
35:    $\eta^* \leftarrow \arg \max_{\eta \in \mathcal{M}} \Delta h_\eta$  ▷ best from shared memory
36:    $\mathcal{H}_{\text{sim}} \leftarrow \text{SimilaritySample}(\mathcal{M}, f_t)$  ▷ embedding-based retrieval
37:    $\mathcal{H}_{\text{cand}} \leftarrow \{\eta_c\}_{c \in \mathcal{C}_t} \cup \{\eta^*\} \cup \mathcal{H}_{\text{sim}}$ 
38:    $\eta_{t+1} \leftarrow \text{LLMSelect}(\mathcal{H}_{\text{cand}}, \{f_\eta\}_{\eta \in \mathcal{M}})$  ▷ select / modify / generate
39: end if
40:
41: ▷ Stage 6: Solution Update
42:  $\tau \leftarrow \text{Sketch}(\eta_{t+1}, s^*)$  ▷ code modification plan
43:  $s_{t+1} \leftarrow \text{Implement}(\tau, s^*)$  ▷ iterative refinement until executable
44: return  $(\eta_{t+1}, s_{t+1})$ 
```

- $\{\eta_c\}_{c \in C_t^{(i)}}$: Hypotheses generated by the current trace based on local challenges (§3.4)
- η^* : The hypothesis from the highest-scoring iteration in the shared success memory \mathcal{M}
- \mathcal{H}_{sim} : Hypotheses sampled via the probabilistic interaction kernel

Probabilistic Interaction Kernel. Inspired by statistical physics (Isihara, 2013), we introduce a probabilistic interaction mechanism that enables controlled information exchange between traces. For each hypothesis η_c generated by the current trace, we compute an interaction potential with all hypotheses η_j in the shared memory \mathcal{M} :

$$U_{cj} = \alpha \cdot S_{cj} \cdot e^{-\gamma L} + \beta \cdot \tanh(\Delta_{cj}) \in [-2, 2], \quad (14)$$

where U_{cj} is the interaction potential between hypothesis η_c and historical hypothesis η_j . The parameters α and β are weights controlling the relative importance of embedding similarity S_{cj} (cosine similarity between embeddings of η_c and η_j) and score difference Δ_{cj} . The parameter γ is a decay factor based on the iteration count L .

The score difference Δ_{cj} is defined as:

$$\Delta_{cj} = \begin{cases} h_j - h^{*(i)}, & \text{if higher score is better} \\ h^{*(i)} - h_j, & \text{if lower score is better} \end{cases} \quad (15)$$

where $h^{*(i)}$ is the trace-local best score and h_j is the score associated with hypothesis η_j . The sampling distribution is computed via softmax normalization:

$$p_{cj} = \frac{\exp(U_{cj})}{\sum_k \exp(U_{ck})}, \quad \eta_{\text{sim}} \sim \text{Categorical}(p_{cj}). \quad (16)$$

This interaction potential integrates both semantic information from hypothesis text and empirical information from scores. The decay factor $e^{-\gamma L}$ applied to the similarity term reflects that the optimization trajectory is not a Markov process: the generation of later hypotheses depends on multiple previous steps. Therefore, in later stages of exploration, the weight of this component decays rapidly, so that score information plays a more dominant role, biasing selection toward empirically proven strategies.

C.3.2 Adaptive Hypothesis Selection

After constructing the candidate pool, an LLM-based selector produces the final hypothesis. The selector is not constrained to choose from the candidates; instead, it can take one of three actions:

- **Select:** Choose the most promising hypothesis from $\mathcal{H}_{\text{cand}}$
- **Modify:** Revise an existing candidate to improve it (e.g., adjust hyperparameters, refine approach)
- **Generate:** Create a new hypothesis by synthesizing insights from multiple candidates

This design aims to reduce hallucinations and stabilize outcomes across different traces, as the provided candidates serve as grounded references rather than strict constraints.

Algorithm 3 presents the cross-trace hypothesis selection process.

This adaptive selection mechanism, combined with the probabilistic interaction kernel, enables efficient cross-trace learning without sacrificing exploration diversity. When a trace falls behind the global best, it can leverage successful strategies from other traces via η^* and \mathcal{H}_{sim} ; when leading, it continues local exploration while still benefiting from the shared memory. This design ensures that the system leverages collective discoveries from all parallel traces while maintaining independent optimization trajectories.

C.4 Algorithm for GOME Main Loop

Algorithm 4 orchestrates the overall optimization process. It begins with “Initialization with Forced Diversification”, where starting hypotheses are sequentially conditioned on prior proposals to maximize the initial coverage of the solution space.

The core “Parallel Optimization” then coordinates multiple concurrent traces. In each iteration, the system: (1) invokes the feedback mechanism (Algorithm 1) to evaluate the current state; (2) updates the Global Success Memory \mathcal{M} upon acceptance, allowing successful strategies to be shared across traces; and (3) calls the structured reasoning module (Algorithm 2) to generate the next solution. The process concludes with Final Selection, which employs multi-seed evaluation on top candidates to mitigate variance and ensure the robustness of the final output s^* .

D Qualitative Analysis and Case Studies

We provide detailed case studies to illustrate GOME’s reasoning capabilities, contrasting it with scalar-driven baselines and demonstrating its performance in real-world scenarios.

Algorithm 3 Cross-trace Hypothesis Selection

```
1: Input: local hypotheses  $\{\eta_c\}_{c \in \mathcal{C}_t^{(i)}}$ , success memory  $\mathcal{M}$ , trace-local best score  $h^{*(i)}$ 
2: Output: selected hypothesis  $\eta_{t+1}^{(i)}$ 
3:
4:
5:  $\eta^* \leftarrow \arg \max_{\eta \in \mathcal{M}} \Delta h_\eta$  ▷ Step 1: Construct Candidate Pool (§3.5)
6:  $\mathcal{H}_{\text{sim}} \leftarrow \text{SampleByKernel}(\mathcal{M}, \{\eta_c\})$  ▷ best from shared memory
7:  $\mathcal{H}_{\text{cand}} \leftarrow \{\eta_c\}_{c \in \mathcal{C}_t^{(i)}} \cup \{\eta^*\} \cup \mathcal{H}_{\text{sim}}$  ▷ probabilistic interaction
8:
9:
10:  $\eta_{t+1}^{(i)} \leftarrow \text{LLMSelect}(\mathcal{H}_{\text{cand}}, \{f_\eta\}_{\eta \in \mathcal{M}})$  ▷ Step 2: LLM-based Selection
11: ▷ Selector can: Select from candidates, Modify existing, or Generate new
12: return  $\eta_{t+1}^{(i)}$ 
```

Algorithm 4 GOME: Gradient-based Optimization for Machine Learning Engineering

```
1: Input: task specification  $\mathcal{T}$ , dataset  $\mathcal{D}$ , number of traces  $N$ , time budget  $B$ 
2: Output: best solution  $s^*$ 
3:
4:
5:  $\mathcal{M} \leftarrow \emptyset$  ▷ Phase 1: Initialization with Forced Diversification (§3.5)
6: for  $n = 1$  to  $N$  do ▷ global success memory
7:    $\mathcal{H}^{(n)} \leftarrow \emptyset$  ▷ local history for trace  $n$ 
8:    $\eta_0^{(n)} \leftarrow \text{GenerateInitialHypothesis}(\mathcal{T}, \{\eta_0^{(j)}\}_{j < n})$  ▷ condition on earlier proposals
9:    $s_0^{(n)} \leftarrow \text{Implement}(\eta_0^{(n)}, \mathcal{T})$  ▷ initial solution for trace  $n$ 
10:   $s^{*(n)} \leftarrow s_0^{(n)}$  ▷ trace-local best
11: end for
12:
13:
14: while  $\text{RemainingBudget}(B) > 0$  do ▷ Phase 2: Parallel Optimization
15:   for each trace  $n \in \{1, \dots, N\}$  in parallel do
16:
17:   ▷ Step 1: Execute and Validate (Algorithm 1)
18:    $(f_t^{(n)}, d_t^{(n)}) \leftarrow \text{FeedbackAndValidation}(s_t^{(n)}, s^{*(n)}, \mathcal{T})$ 
19:
20:   ▷ Step 2: Update Success Memory (§3.3)
21:   if  $d_t^{(n)} = 1$  then ▷ accepted
22:      $\Delta h_t^{(n)} \leftarrow h_t^{(n)} - h^{*(n)}$ 
23:      $\mathcal{M} \leftarrow \mathcal{M} \cup \{(\eta_t^{(n)}, f_t^{(n)}, \Delta h_t^{(n)})\}$ 
24:     if  $h_t^{(n)} > h^{*(n)}$  then
25:        $s^{*(n)} \leftarrow s_t^{(n)}$  ▷ update trace-local best
26:     end if
27:   end if
28:    $\mathcal{H}^{(n)} \leftarrow \mathcal{H}^{(n)} \cup \{(s_t^{(n)}, f_t^{(n)}, d_t^{(n)})\}$ 
29:
30:   ▷ Step 3: Structured Reasoning (Algorithm 2)
31:    $(\eta_{t+1}^{(n)}, s_{t+1}^{(n)}) \leftarrow \text{StructuredReasoning}(f_t^{(n)}, \mathcal{H}^{(n)}, \mathcal{M}, s^{*(n)}, \mathcal{T})$ 
32: end for
33:
34: ▷ Adaptive Time Management (§3.6)
35:    $B \leftarrow \text{AdjustBudget}(B, \mathcal{M}, \{f_t^{(n)}\}_{n=1}^N)$ 
36: end while
37:
38: ▷ Phase 3: Final Selection (§3.6)
39:  $\mathcal{S}_{\text{top}} \leftarrow \text{TopK}(\{s^{*(n)}\}_{n=1}^N, k)$ 
40: for  $s \in \mathcal{S}_{\text{top}}$  do
41:    $\bar{h}(s) \leftarrow \text{MultiSeedEval}(s, \mathcal{T})$  ▷ reduce variance
42: end for
43:  $s^* \leftarrow \arg \max_{s \in \mathcal{S}_{\text{top}}} \bar{h}(s)$ 
44: return  $s^*$ 
```

D.1 Preventing Overfitting via Hierarchical Validation

A critical limitation of score-centric feedback utilization is the reliance on validation scores as the sole proxy for solution quality. This often leads to overfitting or the acceptance of technically flawed solutions that happen to score well on a specific split.

We analyze GOME’s overfitting detection capability on the Stanford COVID Vaccine task from MLE-Bench (High tier), where the goal is to predict RNA degradation rates (MCRMSE, lower is better). Across 90 optimization iterations, we identified 9 cases where validation improved but test performance degraded, which is the hallmark of overfitting.

Detection Performance. Table 13 summarizes GOME’s overfitting detection results. The hierarchical validation correctly rejected 6 out of 9 overfitting cases (66.7% detection rate), preventing potentially harmful updates from being committed.

Metric	Value
Total optimization iterations	90
Overfitting cases identified	9
Correctly rejected	6 (66.7%)
Incorrectly accepted	3 (33.3%)

Table 13: Overfitting detection performance on Stanford COVID Vaccine.

Case Study: Loop 32. Figure 6 presents a representative case where GOME correctly rejected a deceptive improvement. The agent proposed a loss reweighting strategy that showed dramatic validation improvement (-57.6%) but would have caused catastrophic test degradation ($+137.5\%$).

Detailed Overfitting Cases. Table 14 lists all identified overfitting cases and GOME’s decisions. The three incorrectly accepted cases (Nodes 40, 49, 80) all exhibited marginal validation improvements ($< 6\%$) with small test regressions ($< 6.3\%$), making them harder to distinguish from noise. In contrast, the correctly rejected cases often showed larger validation-test divergence patterns.

Analysis. This case demonstrates why methods that use feedback primarily for scalar ranking are fundamentally limited: they cannot distinguish genuine improvements from deceptive ones. A score-driven agent would have accepted Loop 32’s

Loop	Component	Δh_{val}	Δh_{test}	Decision	Correct
23	Model	-63.8%	$+0.3\%$	Reject	✓
26	FeatureEng	-5.1%	$+1.4\%$	Reject	✓
32	Model	-57.6%	$+137.5\%$	Reject	✓
40	Model	-0.2%	$+6.3\%$	Accept	✗
47	Ensemble	-11.9%	$+0.4\%$	Reject	✓
49	Model	-5.8%	$+1.2\%$	Accept	✗
57	Model	-69.8%	$+0.3\%$	Reject	✓
77	FeatureEng	-0.1%	$+0.03\%$	Reject	✓
80	Model	-0.1%	$+0.03\%$	Accept	✗

Table 14: All overfitting cases in Stanford COVID Vaccine. Δh_{val} : validation change (negative is improvement). Δh_{test} : test change (positive is degradation).

solution based on the 57.6% validation “improvement,” leading to catastrophic test degradation. GOME’s structured feedback, which analyzes code changes, training dynamics, and feature correlations, provides the semantic information necessary to detect such patterns. The 66.7% detection rate, while not perfect, represents a significant advantage over purely score-driven approaches that would accept all 9 overfitting cases (0% detection).

D.2 Cross-Task Analysis of Deceptive Improvements

Our earlier analysis focused on a single representative task. This section extends the analysis across multiple MLE-Bench tasks to identify systematic patterns in deceptive improvements—iterations where validation metrics improve but test performance degrades.

Task-Type Vulnerability. Examining GPT-5 optimization traces across the full MLE-Bench, we find that deceptive improvements are not uniformly distributed but concentrate in two task categories:

- High-dimensional sparse-feature tasks.** Tasks such as RNA sequence modeling (stanford-covid-vaccine), genomic prediction, and high-cardinality categorical problems are most vulnerable. The high dimensionality and feature sparsity create opportunities for complex loss reweighting or subtle feature engineering to exploit spurious correlations that appear genuine on limited validation splits but fail to transfer to the test distribution.
- Small-sample time-series tasks.** Tasks with limited temporal data are prone to temporal feature engineering that overfits to idiosyncratic patterns in the training window. Highly specific lag features, holiday indicators, or trend decompositions can capture noise rather than signal, producing validation gains that do not transfer across the temporal boundary to

Case: Overfitting Detection on Stanford COVID Vaccine (Node 32)

Task: Predict RNA degradation rates (MCRMSE, lower is better)

Hypothesis η_t : Align training objective with evaluation by redefining loss weights: for reactivity use clipped inverse-variance weights $w = \text{clip}(1/(\text{err}^2 + 10^{-6}), 0.1, 10)$; for `deg_Mg_pH10` and `deg_Mg_50C` use constant $w = 1.0$. Replace hard `SN_filter` gating with soft factor (1.0 for `SN=1`, 0.5 for `SN=0`). Per-batch renormalize so each scored channel has equal total weight mass.

Metrics:

- Previous (Parent 27): $h_{\text{val}}^{*(i)} = 0.6948$, $h_{\text{test}}^{*(i)} = 0.2688$
- Current (Node 32): $h_{t,\text{val}} = 0.2945$ (\downarrow 57.6% better), $h_{t,\text{test}} = 0.6382$ (\uparrow 137.5% worse)

Code Changes: Kept PairAwareBiGRU architecture, features, split, inference identical to SOTA. Replaced training weights with channel-specific weighting; introduced per-batch channel renormalization; left validation metric as unweighted MCRMSE.

Observations from f_t : Current MCRMSE is 0.2945 vs. SOTA 0.2559; performance degraded by +0.0385. EDA indicates weak target–error association for reactivity (Pearson ≈ -0.042 , Spearman ≈ 0.280), suggesting inverse-variance weighting may not benefit learning.

Hypothesis Evaluation: REFUTED. The redesigned loss weighting and per-batch renormalization worsened validation MCRMSE. Likely causes: weak correlation between reactivity and its error; soft SN scaling introducing noisier samples; renormalization potentially overemphasizing harder channels.

Decision: $d_t = 0$ (**REJECT**) — Despite validation appearing to improve, the structured analysis identified that the “improvement” stemmed from metric misalignment rather than genuine generalization gains.

Figure 6: Overfitting detection on Stanford COVID Vaccine. The validation score appeared to improve by 57.6%, but GOME’s hierarchical validation analyzed the code changes and training dynamics to correctly reject the solution, avoiding a catastrophic 137.5% test regression.

the test period.

Modification Taxonomy. Across the identified deceptive improvements, three dominant modification patterns emerge:

1. **Loss reweighting and objective misalignment.** The agent modifies the training loss (e.g., channel-specific weighting, sample reweighting, focal loss variants) in ways that align well with the validation metric on the current split but introduce systematic bias. This category is the most reliably detected by hierarchical validation, as the code changes concentrate in the training loop, making it relatively straightforward for the LLM validator to identify discrepancies between training objective and evaluation metric.
2. **Aggressive feature engineering.** The agent introduces highly specific features—interaction terms between rare categories, narrow time windows, task-specific hand-crafted indicators—that capture noise in the training/validation split but encode distributional artifacts rather than causal patterns. These are moderately detectable, as the validator can flag suspiciously specific features by analyzing their construction logic.
3. **Simplicity Bias.** After repeated failed itera-

tions, the agent sometimes resorts to surface-level heuristics or hard-coded shortcut rules: threshold-based classifications mapping input ranges directly to output labels, median imputation strategies that “hack” the evaluation metric without learning the underlying data logic, or constant-prediction fallbacks exploiting metric asymmetries. This pattern aligns with the “shortcut solvability” phenomenon (Nie et al., 2026), where agents prioritize superficial metric gains over genuine reasoning. Simplicity Bias is the hardest category to detect, because shortcut solutions often produce clean, syntactically reasonable code—the validator must reason about the *semantic adequacy* of the approach relative to the task complexity, which requires deeper understanding.

D.3 Comparison with MLE-STAR

MLE-STAR (Nam et al., 2025) represents a recent advancement in MLE agents, achieving 63.6% medal rate on MLE-Bench Lite through web retrieval and targeted code-block refinement. While MLE-STAR and GOME share a chain-based structure (as opposed to tree or graph topologies), they differ fundamentally in how feedback guides optimization. We analyze these differences across three

dimensions.

Feedback Utilization: Ranking vs. Update.

MLE-STAR employs a two-level selection process: (1) an ablation study identifies which code block has the greatest performance impact, and (2) multiple refinement plans are generated and evaluated, with the highest-scoring plan selected. Formally, given K candidate plans $\{p_k\}_{k=0}^{K-1}$ for code block c_t , MLE-STAR selects $k^* = \arg \max_k h(s_t^k)$ where $h(\cdot)$ is the validation score. This is fundamentally a ranking operation: scalar scores determine which pre-generated candidate survives.

In contrast, GOME uses structured reasoning to generate directional updates. Rather than generating multiple candidates and selecting via scores, the LLM analyzes execution feedback to produce a single improvement hypothesis with explicit direction (what to change) and magnitude (how much). The validation score serves as a stopping criterion and acceptance gate, not a selection mechanism among alternatives.

Code Block Selection: Ablation vs. Reasoning.

MLE-STAR’s ablation study agent \mathcal{A}_{abl} generates code that systematically disables or modifies components (typically 2–3 per iteration), then compares resulting scores to identify the most impactful block. This requires additional exploratory executions per iteration and provides only which block matters most, not why it underperforms or how to improve it.

GOME’s structured reasoning operates on execution feedback directly: training dynamics, validation curves, error patterns, and code behavior. The LLM reasons about why current performance is limited and what modification would address the underlying issue. This reasoning-first approach generates actionable hypotheses without requiring additional ablation executions.

Illustrative Comparison. Consider the overfitting case from §D.1 where validation improved 57.6% but test degraded 137.5%. This represents subtle overfitting that does not involve explicit data leakage detectable by rule-based checkers. Under MLE-STAR’s framework, once the data leakage checker passes, the ablation study would identify the model component as impactful based on score changes. The planner would then generate multiple refinement plans, and plan selection would rely solely on validation scores, ultimately accepting the overfitting solution due to its apparent 57.6%

improvement.

Under GOME’s framework, structured reasoning analyzes training dynamics, feature correlations, and the semantic nature of the code change. This analysis identifies weak target-error correlation (Pearson ≈ -0.042), recognizing that the validation improvement stems from spurious fitting rather than genuine generalization. The hierarchical validation gate consequently rejects the update based on this reasoning rather than the misleading validation score.

Summary. Table 15 summarizes the key distinctions. While MLE-STAR advances beyond AIDE’s tree search through targeted refinement, it remains within the score-driven paradigm: validation scores ultimately determine which solutions propagate. GOME shifts to reasoning-driven optimization, where LLM analysis of execution signals generates improvement direction rather than merely selecting among candidates. This distinction becomes increasingly important as model reasoning capabilities improve, enabling more accurate “gradient” estimation.

Aspect	MLE-STAR	GOME
Feedback role	Ranking	Update
Plan generation	Multiple candidates	Single hypothesis
Selection mechanism	arg max score	Reasoning gate
Block identification	Ablation study	Structured analysis
Subtle overfitting detection	0% (score-driven)	66.7% (reasoning)

Table 15: Key distinctions between MLE-STAR and GOME. Overfitting detection rate is based on the case study in §D.1, where score-driven methods would accept all 9 overfitting cases.

D.4 Online Kaggle Validation: Store Sales Forecasting

While MLE-Bench provides standardized evaluation, real-world data science requires navigating uncurated data, ambiguous constraints, and complex domain logic. To validate GOME’s effectiveness in this setting, we deployed the framework on an active Kaggle competition: Store Sales - Time Series Forecasting.⁷ This competition requires forecasting unit sales for thousands of product families sold at Corporación Favorita stores in Ecuador over a 15-day horizon. The dataset includes store metadata, oil prices (Ecuador’s economy is oil-dependent), promotional information, and a calendar of holidays with complex transfer rules. The evaluation metric

⁷<https://www.kaggle.com/competitions/store-sales-time-series-forecasting>

is Root Mean Squared Logarithmic Error (RMSLE, lower is better). GOME achieved a final leaderboard RMSLE of 0.431, ranking in the **top 15%** of all participants.

Optimization Trajectory. Table 16 summarizes the optimization process. Starting from an initial baseline (validation RMSLE 0.485), GOME progressively refined the solution over 48 iterations. The best validation score (0.318) was achieved at Loop 37, which translated to a leaderboard RMSLE of 0.431, outperforming approximately 85% of human participants.

Loop	Val RMSLE	Key Modification	Outcome
0	0.485	Initial baseline	Accepted
2	0.354	Two-stage hurdle architecture	Accepted
7	0.327	Rolling statistics & lag features	Accepted
28	0.318	Store-local holiday processing	Accepted
37	0.318	Horizon bucketing & calibration	Accepted (best)
30	(0.547)	Naive ensemble strategy	Rejected
43	(0.559)	Unstable pipeline modification	Rejected
–	0.378	Human best	–

Table 16: **Optimization trajectory on the Store Sales competition.** All intermediate scores are internal validation RMSLE (15-day holdout). The trajectory shows steady improvement from 0.485 to 0.318 over 37 iterations, with the gate mechanism rejecting degrading modifications (Loops 30, 43). The final submission achieved a leaderboard RMSLE of 0.431. Note that the human leaderboard leader (0.378) employs multi-level blending of other participants’ submissions.

This trajectory demonstrates GOME’s effective gate mechanism. At Loop 30, an ensemble strategy caused severe metric degradation (0.318→0.547), which was correctly rejected. Similarly, Loop 43’s workflow modification destabilized the pipeline (0.559) and was rejected. The system preserved the robust solution from Loop 37 throughout subsequent iterations.

Architectural Comparison. To contextualize GOME’s performance, we analyzed the codebases of high-ranking public notebooks. As shown in Figure 7, a striking distinction emerges. Many top-tier public solutions employ second-level blending—scripts that optimally weight prediction files submitted by other competitors using hand-tuned coefficients and rank-based corrections. These function as meta-optimization tools for existing outputs rather than building models from raw data.

In contrast, GOME functions as an autonomous data scientist, constructing the entire pipeline from raw relational tables: (1) performing extensive feature engineering including horizon bucketing,

rolling statistics, and store-local holiday processing; (2) training two-stage hurdle models (classifier for zero/non-zero, regressor for magnitude) with recency-weighted sampling; and (3) implementing pooled Platt calibration for probability adjustment. This confirms that GOME achieves expert-level performance through genuine architectural reasoning rather than aggregation of existing solutions.

D.5 Error Analysis

To understand GOME’s limitations, we conduct a detailed case study on the stanford-covid-vaccine task (a high-complexity task in MLE-Bench) using GPT-5 as the backbone model.

Category	Count	Percentage
Successful Updates	31	34.4%
Gradient Hallucination	35	38.9%
Resource Constraints	21	23.3%
Validation Blindspots	3	3.3%
Total Failures	59	65.6%

Table 17: Failure mode distribution on stanford-covid-vaccine (90 iterations, GPT-5).

Table 17 summarizes the failure mode distribution across 90 total iterations. We classify unsuccessful iterations into four categories: (1) **Gradient Hallucination**, where the reasoning module produces confident but incorrect improvement directions; (2) **Implementation Failures**, where correct hypotheses fail during code generation due to API misuse or syntax errors; (3) **Resource Constraints**, where solutions exceed memory limits or time budget B ; and (4) **Validation Blindspots**, where validation score improves but test performance degrades.

Gradient hallucination dominates at 38.9%, where the reasoning module generates plausible but incorrect optimization directions. This rate is expected to decrease as LLM reasoning capabilities improve, consistent with our scaling claim (§5). Resource constraints account for 23.3% of failures due to hardware limitations (single V100 GPU, 32GB memory) and time budget restrictions. Many proposed solutions exceeded available memory or the time budget, particularly for this RNA sequence modeling task with computationally intensive architectures. Allocating more powerful hardware or extending time budgets would likely reduce this failure mode. Validation blindspots (3.3%) are rarer but harder to address, as they evade the gate mecha-

(a) Top-Tier Human Solution: Post-Hoc Blending

```
# 1. Ingest pre-calculated prediction files from other participants (No training)
def read(dk, i):
    FiN = dk["path"] + dk["subj"] + dk["name"] + ".csv" # e.g., "0.37982.csv"
    return pd.read_csv(FiN)
# 2. Hard-coded weights for linear combination (manually tuned)
params = {
    'subwts': [+14, -1, -5, -8] / 100, # correction weights by rank
    'subj': [{'name': '0.37982', 'weight': 0.30}, {'name': '0.37984', 'weight': 0.10},
             {'name': '0.38006', 'weight': 0.30}, {'name': '0.38040', 'weight': 0.30}]
}
# 3. Compute weighted sum (Pure ensembling of existing outputs)
def correct(x, cols, weights, subwts):
    rank_indices = [x['alls'].index(c) for c in cols] # rank-based reweighting
    return sum([x[cols[j]] * (weights[j] + subwts[rank_indices[j]]) for j in range(len(cols))])
```

(b) GOME Solution: Ab Initio Modeling

```
# 1. Feature Engineering from Raw Data (Reasoning-driven)
buckets = {'H1': [1,2,3], 'H2': [4,5,6,7], 'H3': [8,9,10,11,12,13,14,15]}
# 2. Two-Stage Hurdle Model Training (Learning from scratch)
for bname, horizons in buckets.items():
    clf = LGBMClassifier(n_estimators=4000, num_leaves=384, learning_rate=0.03)
    clf.fit(X_clf_tr, y_clf_tr, sample_weight=w_recency)

    reg = LGBMRegressor(n_estimators=6000, num_leaves=384, learning_rate=0.03)
    reg.fit(X_reg_tr, y_reg_tr, sample_weight=w_recency)

    calibrator = PooledPlattCalibrator(max_h=max(horizons))
    calibrator.fit(clf.predict_proba(X_val)[:,-1], h_val, y_val)
# 3. Model Inference: P(nonzero) * E[sales | nonzero]
yhat = calibrator.transform(clf.predict_proba(X)[:,-1], h) * np.expml(reg.predict(X))
```

Figure 7: **Codebase comparison.** (a) High-ranking public notebooks often use blending scripts that aggregate prediction CSVs submitted by other participants, a form of meta-optimization that, while effective for leaderboard climbing, involves no model training. GOME constructs the full pipeline from raw relational tables under the closed-world protocol without any external knowledge, including feature engineering, horizon-bucketed model training, and Platt calibration.

nism entirely: the proposed modification improves validation metrics but degrades test performance due to distribution shifts. More sophisticated validation strategies may help mitigate this issue in future work.

E Theoretical Motivation for the Smoothness Assumption

GOME’s gradient-based paradigm assumes that the MLE optimization landscape is locally smooth enough to support directed, incremental refinement. This section summarizes the theoretical rationale and empirical evidence for this assumption.

Loss-Surface Connectivity in Neural Networks.

A growing line of optimization research indicates that neural-network loss surfaces are substantially more benign than worst-case non-convex intuition implies. Garipov et al. (2018) show that independently trained optima can often be connected by simple low-loss paths (“mode connectivity”) rather than separated by high-loss barriers. Bottou et al.

(2018) further argue that stochastic gradient methods are effective in overparameterized regimes because optimization trajectories frequently traverse locally smooth, connected regions. Taken together, these results imply that moving from a good solution to a better one often proceeds through a sequence of small improvements rather than discontinuous jumps.

From Parameter Space to Code Space. GOME optimizes code rather than parameters directly. The smoothness intuition transfers to code space because GOME’s structured reasoning does not propose arbitrary syntactic edits; instead, it uses diagnostics (training dynamics, validation curves, error distributions, and feature-importance patterns) to target changes in the *training procedure*—for example, learning-rate schedules, feature engineering, loss design, or regularization. These behavior-level modifications induce controlled changes in the underlying optimization trajectory. In this sense, GOME’s “gradient” operates on the pipeline’s *functional behavior*, not raw syntax. This is a

weaker requirement than global smoothness of the code-to-performance mapping: it only requires that *reasoning-guided* edits produce locally predictable performance changes.

Empirical Evidence. Our experiments provide practical evidence for local smoothness:

- **High per-iteration success rate.** Table 5 reports a 41.1% improvement rate, meaning that nearly half of optimization steps yield validated gains through incremental updates. If the landscape were dominated by discontinuities or isolated basins, this rate would be much lower.
- **Sustained convergence trajectories.** Figure 4 shows smooth, steep descent for GOME on Frontier-tier models, consistent with gradient-style optimization in locally smooth regions. By contrast, MCTS exhibits more stepwise, plateau-heavy dynamics that are typical of search in rougher landscapes.
- **Kaggle case study.** In the Store Sales deployment (Appendix D.4), performance improves progressively over 48 iterations (RM-SLE 0.485 \rightarrow 0.318), with degrading modifications filtered out by the gate mechanism. This monotonic trajectory is consistent with iterative optimization along a smooth local path.

Handling Non-Smooth Transitions. Some transitions in MLE are intrinsically discontinuous (e.g., switching from linear models to neural networks, or from single models to ensembles). GOME mitigates this with two mechanisms:

1. **Multi-trace forced diversification** (§3.5) initializes N traces with distinct architectural hypotheses (e.g., gradient boosting vs. neural networks vs. ensembles), providing broad coverage of separate regions in solution space. This is analogous to multi-start optimization in classical non-convex settings (Bottou et al., 2018).
2. **Cross-trace sharing via the probabilistic interaction kernel** (Appendix C.3) allows traces to exchange validated strategies. When one trace discovers a better region, others can adopt or adapt that hypothesis, enabling non-local transitions while preserving local refinement within each region.

Scope. The smoothness assumption is not universal. Tasks with extremely sparse rewards or patho-

logical loss landscapes may violate local smoothness. However, in competitive data-science settings, top solutions often share backbone algorithms and differ mainly in feature engineering, hyperparameter tuning, and ensembling. This structure suggests that the MLE landscape is sufficiently smooth for gradient-based optimization to be effective in practice. The scaling analysis in §5 provides further indirect support: the monotonic widening of GOME’s advantage with model capability is consistent with stronger reasoners producing higher-fidelity gradient signals that better exploit local smoothness.