SCALABLE SKILL RETRIEVAL FOR MACHINE LEARNING AGENTS VIA EVOLVING STATE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning agents solve tasks through sequences of costly experiments, during which their knowledge needs evolve with observed results. Static retrieval, which retrieves skills only once at the beginning of a task, cannot adapt to these changing needs. We study skill retrieval as a sequential decision problem conditioned on experimental progress. Preliminary study on MLE-bench reveal that repeatedly retrieving from the latest experimental state may introduce irrelevant guidance, disrupt promising experiments, and consume resources without improving outcomes. This observation motivates three design goals: retrieving the **right** skill to address the actual problem encountered in the current experiment; retrieving at the **right** time, only when the model's own knowledge is insufficient to solve the current problem; and using the retrieved skill in the **right** way, correctly executing it in the current experimental context. We propose an experiment-state-aware retrieval framework that maintains a structured record of attempted approaches, validation evidence, unresolved failures, and remaining resources. The framework triggers retrieval when this evidence identifies a concrete need, selects skills according to their applicability and feasibility within the remaining budget, and integrates them into targeted experiments with explicit evaluation criteria. Subsequent observations update the experimental state and inform future retrieval decisions. We outline a controlled evaluation on MLE-bench that compares static, naive dynamic, and selective state-aware retrieval under matched resource budgets, measuring final task performance, experimental efficiency, and skill adoption. This work investigates when adaptive access to procedural knowledge can improve long-horizon experimentation and when retrieval instead becomes a source of interference.

1 INTRODUCTION

2 RELATED WORK

2.1 SKILL RETRIEVAL

Skill retrieval for agent has progressed from relevance matching to structured and adaptive selection. SkillRouter (Zheng et al., 2026) uses a retrieve-and-rerank pipeline, while Graph-of-Skills (Liu et al., 2026a) and CaSKG (Li et al., 2026) exploit relationships among skills. SE-GoS (Fu et al., 2026) further updates retrieval structures using execution experience. Retrieval also differs in its timing and conditioning information: SkillsVote (Liu et al., 2026b) performs agentic recommendation before task execution; SkillDAG (Bai et al., 2026) exposes a skill graph that agents can query during execution; and SkillReranker (Chen et al., 2026) derives intermediate task states from task descriptions to guide skill selection.

Existing methods support retrieval during execution and adaptation from historical feedback. For MLE, a further requirement is to determine whether a reusable skill has evidence of benefit under the current experimental conditions and can be executed with the available resources. Semantic relevance, skill dependencies, and prior use alone do not establish whether a procedure addresses the observed failure, can use the available models or predictions, or fits the remaining training budget. Our retrieval design matches applicability conditions and experimental evidence accumulated across

Table 1: **Main results on MLE-bench (full 75 competitions).** We report Any Medal (%) across the Low, Medium, High, and full 75-task splits. Public baselines are copied from the official MLE-bench leaderboard and report mean±SEM over runs.

Agent	Backbone	Time (h)	Low (n=22)	Medium (n=38)	High (n=15)	All (n=75)
Famou-Agent 2.0 (Li et al., 2025a)	Gemini-3-Pro-Preview	24	80.30 ± 1.52	64.04 ± 2.32	42.22 ± 2.22	64.44 ± 1.18
AIBuildAI (Zhang et al., 2026a)	Claude-Opus-4.6	24	77.27 ± 0.00	61.40 ± 0.88	46.67 ± 0.00	63.11 ± 0.44
CAIR MARS+ (Chen et al., 2026c)	Gemini-3-Pro-Preview	24	78.79 ± 1.52	60.53 ± 1.52	44.44 ± 2.22	62.67 ± 0.77
MLEvolve (Du et al., 2026)	Gemini-3-Pro-Preview	12	80.30 ± 1.52	57.89 ± 1.52	42.22 ± 2.22	61.33 ± 1.33
PiEvolve (Botla et al., 2025)	Gemini-3-Pro-Preview	24	80.30 ± 1.52	58.77 ± 0.88	40.00 ± 0.00	61.33 ± 0.77
Thesis (Thesis, 2026)	GPT-5-Codex	24	65.15 ± 1.52	45.61 ± 7.18	31.11 ± 2.22	48.44 ± 3.64
R&D-Agent (Zhang et al., 2026b)	GPT-5	12	68.18 ± 2.62	21.05 ± 1.52	22.22 ± 2.22	35.11 ± 0.44
Codex	GPT-5.5	24	42.42 ± 6.60	31.58 ± 1.52	13.33 ± 3.85	31.11 ± 2.22
Codex + AREX-SKILL	GPT-5.5	48	$\mathbf{86.36 \pm 2.62}$	$\mathbf{69.30 \pm 3.16}$	$\mathbf{62.22 \pm 2.22}$	$\mathbf{72.89 \pm 1.18}$
RRR	GPT5.5	12	90.90 ± 0.00			

tasks to the evolving search state, prioritizing skills that are both supported by relevant evidence and feasible to apply.

3 METHOD

3.1 SKILL REPRESENTATION AND CONSTRUCTION

TBD.

3.2 OFFLINE SKILL EVOLUTION

TBD.

3.3 STATE-CONDITIONED SKILL RETRIEVAL

TBD.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

4.2 MAIN RESULTS

TBD.

4.3 ABLATION STUDIES

TBD.

4.4 CONSTRUCTION, EVOLUTION, AND VALIDATION COSTS

TBD.

5 DISCUSSION

TBD.

6 CONCLUSION

TBD.

AI USE STATEMENT

TBD.

REPRODUCIBILITY STATEMENT

TBD.

REFERENCES

Tong Bai, Zhenglin Wan, Pengfei Zhou, Xingrui Yu, Yang You, and Ivor W. Tsang. SkillDAG: Self-Evolving Typed Skill Graphs for LLM Skill Selection at Scale, 2026. URL https://arxiv.org/abs/2606.03056.

Yanping Chen, Weijie Shi, Wen Yang, and Jiajie Xu. Task Decomposition-Guided Reranking for Adaptive Agent Skill Retrieval, 2026. URL https://arxiv.org/abs/2607.06283.

Dawei Fu, Cheng Jiang, Sitian Qian, Huainan Wang, and Zhongkai Hao. SE-GoS: Self-Evolving Graph-of-Skills for Skill Library at Scale, 2026. URL https://arxiv.org/abs/2609.08228.

Zhiyuan Li, Linyuan Gao, Xuechun Ding, Hongwei Chen, Yuan Wu, and Yi Chang. CaSKG: Counterfactual-Causal Skill Graphs for Scalable Agent Skill Retrieval, 2026. URL https://arxiv.org/abs/2608.25500.

Dawei Liu, Zongxia Li, Hongyang Du, Xiyang Wu, Shihang Gui, Yongbei Kuang, and Lichao Sun. Graph-of-Skills: Dependency-Aware Structural Retrieval for Massive Agent Skills, 2026a. URL https://arxiv.org/abs/2604.05333.

Hongyi Liu, Haoyan Yang, Tao Jiang, Bo Tang, Feiyu Xiong, Yuyu Luo, and Zhiyu Li. SkillsVote: Lifecycle Governance of Agent Skills from Collection, Recommendation to Evolution, 2026b. URL https://arxiv.org/abs/2605.18401.

YanZhao Zheng, ZhenTao Zhang, Chao Ma, YuanQiang Yu, JiHuai Zhu, Yong Wu, Tianze Xu, Baohua Dong, Hangcheng Zhu, Ruohui Huang, and Gang Yu. SkillRouter: Skill Routing for LLM Agents at Scale, 2026. URL https://arxiv.org/abs/2603.22455.

REFERENCES

TBD.

A ADDITIONAL DETAILS

TBD.