SCIENCE CHINA
Information Sciences

• LETTER •

July 2026, Vol. 69, Iss. 7, 179103:1–179103:2
https://doi.org/10.1007/s11432-025-4772-4

ProxyLLM: augmenting LLMs with proxy models for tool utilization through learning from LLMs

Xiaomao ZHOU[1*], Qingmin JIA[1], Yan ZHANG[3], Liwen WANG[3], Renchao XIE[1,2], Tao HUANG[1,2] & Yunjie LIU[1]

[1]*Purple Mountain Laboratories, Nanjing 211111, China*
[2]*School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[3]*China Unicom Research Institute, Beijing 100048, China*

Received 26 May 2025/Revised 15 October 2025/Accepted 23 January 2026/Published online 16 June 2026

Citation Zhou X M, Jia Q M, Zhang Y, et al. ProxyLLM: augmenting LLMs with proxy models for tool utilization through learning from LLMs. Sci China Inf Sci, 2026, 69(7): 179103, https://doi.org/10.1007/s11432-025-4772-4

Despite the promise of tool learning in augmenting large language models (LLMs) with external tools to overcome inherent limitations, e.g., deficits in specialized knowledge, access to real-time information, and nuanced problem-solving ability, current methodologies face significant challenges. Existing approaches [1], primarily based on fine-tuning or in-context learning, are hampered by inflexibility, prompt engineering complexity, and a divergence from real-world assumptions. This study identifies three critical deficiencies in the state-of-the-art: (1) limited tool integration, constrained by LLM token limits and a scarcity of diverse, real-world tools; (2) inferior planning and reasoning, where methods like CoT (chain-of-thought) and ReACT (reasoning and acting) fail to fully elicit model capabilities for complex, multi-step tasks; and (3) low efficiency, a consequence of a single, sequential LLM controller that requires repeated interactions with various tools and lacks parallel processing capabilities. These limitations underscore an urgent need for more effective and efficient strategies that transcend reliance on the LLM's built-in functionalities to achieve robust and scalable tool integration.

To enhance LLM tool utilization, ProxyLLM introduces specialized, lightweight proxy models to improve task planning and tool invocation, thereby boosting accuracy and efficiency. As presented in Figure 1, it delegates the workload from a single LLM to these smaller, domain-specific proxy models. The LLM's role shifts to high-level task decomposition and allocation, using proxy models as intermediaries to select and activate tools, rather than interacting with them directly. This architecture significantly expands the manageable tool scale, reduces errors from long reasoning chains, and improves efficiency. Furthermore, each proxy model enhances tool invocation through skill composition, combining basic tools into more abstract, reusable actions for similar tasks.

Furthermore, to improve the proxy model's ability to learn tool composition, ProxyLLM facilitates the transfer of knowledge from the LLM to the proxy models. This is achieved through the use of LLM-guided deep reinforcement learning (DRL), which significantly boosts the learning efficiency and effectiveness of the proxy models. Specifically, ProxyLLM proposes to transfer the task understanding and decision-making capabilities of LLMs to smaller proxy models in a step-by-step manner. On one hand, it uses data generated by the LLM for knowledge distillation, enabling the proxy models to select tools based on the task. On the other hand, it enhances the decision-making capabilities of the proxy models by guiding DRL with the LLM. Through such an approach, the small proxy models can achieve the same level of task comprehension and logical reasoning abilities as LLMs, enabling them to effectively select and combine different tools.

Transferring tool selection via data distillation. To transfer the tool selection capabilities of LLMs to smaller proxy models, we propose a knowledge distillation framework centered on rationale-based supervision. Our method first leverages the generative power of a teacher LLM to synthesize a specialized dataset of instruction-response pairs, ensuring comprehensive coverage of diverse tool combinations [2]. Crucially, we prompt the LLM, using CoT prompting, to not only select the appropriate tools but also to generate explicit rationales justifying its choices. The proxy model is then trained within a multi-task framework to simultaneously predict the selected tools and reproduce these rationales. By learning the underlying reasoning behind tool selection, the proxy model moves beyond simple imitation to emulate the LLM's complex decision-making process. This rationale-driven supervision yields a more interpretable and effective model, enhancing its ability to generalize and make robust tool selection decisions.

Optimizing tool planning via LLM-guided DRL. To refine tool selection beyond initial proposals, we train a "tool decider" using DRL to choose the optimal tool from top-p candidates at each step. To effectively harness the LLM's expertise, we introduce an expert policy-constrained learning algorithm that aligns the DRL agent with the LLM's high-level decision-making rationale. Since the LLM does not provide direct action probabilities, we first train a Transformer-based policy reconstruction model, F_P, to emulate the LLM's policy. This model, trained via supervised learning on LLM-generated rankings and scores, minimizes the cross-entropy

* Corresponding author (email: xiaomaozhou26@gmail.com)

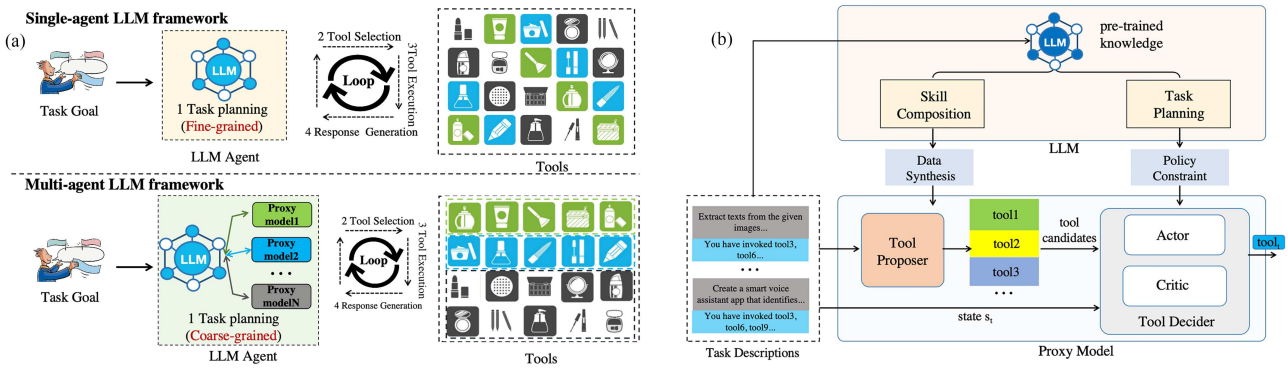

Figure 1 (Color online) The framework of the proposed ProxyLLM. (a) A conceptual comparison of the traditional single-agent LLM framework (top) and the proposed multi-agent LLM framework (bottom); (b) the interaction between the LLM and a proxy model, which consists of a tool proposer and a tool decider.

loss

$$\mathcal{L}_{F_P} = CE(\pi^{LLM}(\cdot|s), F_P(h^z)),$$

where h^z is the output of a self-attention mechanism processing the state embedding.

We then integrate this reconstructed policy into an actor-critic framework with distinct constraints for each network. The learning objective of the DRL agent, thus, becomes a constrained optimization problem, which can be formulated as follows:

$$\begin{aligned} &\max_{\pi} J(\pi) = \max_{\pi} \mathbb{E}_{s_t\sim D, a_t\sim\pi(s_t)}[Q_{\phi_z}(s_t, a_t)] \\ &\text{s.t.} \quad \mathbb{E}(D_{JS}(\pi^{DRL}(\cdot|s_t), \pi^{LLM}(\cdot|s_t))) < \sigma, \end{aligned} \tag{1}$$

where D_{JS} signifies the JS divergence, which is a symmetrized and normalized variant of the Kullback-Leibler (KL) divergence [3], offering a bounded measure of similarity between two probability distributions, ranging from 0 to 1. σ represents the maximum policy deviation.

The actor's objective is to maximize cumulative rewards while maintaining a probability distribution close to the LLM's, a constraint enforced using Jensen-Shannon (JS) divergence. The optimal policy is found by solving

$$\begin{aligned} \pi_\theta^* &= \operatorname*{argmax}_{\pi_\theta} J(\pi_\theta) \\ &= \operatorname*{argmax}_{\pi_\theta} \mathbb{E}_{s_t\sim D, a_t\sim\pi_\theta(s_t)} \Bigg[\sum_{t=0}^{T} \gamma^t r(s_t, a_t) \\ &\quad - \lambda_1 D_{JS}(\pi^{DRL}(\cdot|s_t), \pi^{LLM}(\cdot|s_t)) \Bigg], \end{aligned} \tag{2}$$

where D represents the experience replay buffer, and λ_1 denotes the weight coefficient for the JS constraint.

For the critic, which estimates the action-value function $Q(s, a)$, we align its ranking of tool candidates with the LLM's using Kendall's Tau coefficient (τ_b) [4]. This ensures the critic's value assessments respect the LLM's ordinal preferences. The critic learns by minimizing a modified Bellman error that incorporates this ranking constraint:

$$\begin{aligned} L(\phi_z) =& \mathbb{E}_{(s_t, a_t, r_t, s_{t+1})\sim D}[(Q_{\phi_z}(s_t, a+t) \\ &- (r_t + \gamma V_g(s_{t+1})))^2] - \lambda_2 \tau_b, \end{aligned} \tag{3}$$

where λ_2 denotes the weight coefficient for the Kendal's Tau constraint, and V_g represents a periodically updated target value function.

Experiments. To evaluate ProxyLLM, we developed three new datasets for complex, long-term planning tasks: ToolBench+ (267 tasks) and ToolAlpaca+ (312 tasks), and CPND, a real-world dataset from a computing power network (CPN) with 56 records involving over 60 components across 5 service stages. As detailed in Appendix A, experimental results show that Multi-LLM architectures significantly outperform Single-LLM methods on complex tasks by decomposing them into specialized components. The ProxyLLM framework excels through its task-based division, which enhances accuracy, flexibility, and scalability. Ablation studies confirm that ProxyLLM's efficiency stems from its specialized architecture, enabling cost-effective, high-performance agents by using smaller proxy models.

Conclusion. We introduce ProxyLLM, a novel framework to enhance the tool utilization capabilities of LLMs. It addresses the inefficiency and high error rates of direct LLM invocation by delegating tool selection to an ensemble of smaller, specialized proxy models, thereby improving task accuracy and efficiency. The framework employs a two-step knowledge transfer: first, knowledge distillation using LLM-generated data with rationales to ensure task comprehension; second, LLM-guided DRL to align the proxy models' strategies with the LLM's preferences. Extensive experiments demonstrate that ProxyLLM significantly outperforms existing methods in both task accuracy and invocation efficiency, presenting a promising solution for building generalizable, large-scale intelligent agents.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 92367104, 92267301).

Supporting information Appendix A. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Long S, Tan J, Mao B, et al. A survey on intelligent network operations and performance optimization based on large language models. IEEE Commun Surv Tut, 2025, 27: 3915–3949
2 Kaur S, Park S, Goyal A, et al. Instruct-skillmix: a powerful pipeline for LLM instruction tuning, 2024. ArXiv:2408.14774
3 Majtey A P, Lamberti P W, Prato D P. Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. Phys Rev A, 2005, 72: 052310
4 El-Hashash E F, Shiekh R H A. A comparison of the Pearson, Spearman rank and Kendall Tau correlation coefficients using quantitative variables. Asian J Probab Stat, 2022, 20: 36–48