Predicting Emergent Tool Use in LLMs Before It Emerges: A Proxy Perspective

Bo-Wen Zhang[1], **Yan Yan**[2*], **Guang Liu**[3], **Xu-Cheng Yin**[1]

[1]Univeristy of Science and Technology Beijing, Beijing China
[2]China University of Mining Technology Beijing, Beijing China
[3]Beijing Academy of Artificial Intelligence, Beijing China
bowenzhang@ustb.edu.cn, yanyan@cumtb.edu.cn

Abstract

Tool-use capabilities fundamentally transform large language models (LLMs) from passive language generators into active agents with real-world utility, thus drawing intense research focus. However, as a canonical emergent ability characterized by abrupt onset during training, tool-use defies prediction by conventional scaling laws, hindering principled model design and efficient training. In this work, we propose a proxy-task framework to predict emergent tool-use capabilities by measuring early model performance on carefully selected non-emergent tasks. We quantify each proxy task by two properties: *alignment*, reflecting its correlation with tool-use performance, and *consistency*, indicating stability across diverse training conditions. These metrics guide a weighted aggregation of proxy signals to predict final tool-use rankings. Theoretically, we formalize how such weighted signals approximate emergent tool use under relaxed assumptions with bounded extrapolation guarantees. Empirically, our approach is validated across training checkpoints, model scales, and data setups. Results demonstrate that a properly weighted ensemble of proxy tasks accurately predicts downstream tool-use ability long before it manifests. Our findings provide new theoretical foundations and practical tools for efficient training and capability planning, advancing understanding of emergent behaviors in LLMs.

Introduction

Tool-use capabilities fundamentally transform large language models (LLMs) from passive language generators into dynamic agents capable of interacting with external environments and performing complex, multi-step reasoning. This pivotal ability has attracted intense research interest because it extends the utility of LLMs beyond text generation towards real-world applications such as automated programming, robotic control, and interactive AI systems (Li et al. 2023; Tang et al. 2023; Luo et al. 2025).

Despite its importance, tool-use exemplifies an emergent ability that manifests through a discrete phase transition: no capability is observed in early training, yet substantial competence appears abruptly once models surpass critical thresholds in scale, data, or optimization dynamics. This nonlinear emergence defies conventional scaling laws, which assume smooth and continuous performance improvements and cannot reliably forecast LLMs' final tool-use proficiency from early training performance (Wei 2022; Kaplan et al. 2020). Consequently, early-stage model evaluations under different configurations provide limited guidance for optimizing design and training strategies. To address this challenge, we adopt a **proxy perspective** to investigate the existence of **proxy tasks whose early-stage performance signals reliably correlate with and predict ultimate tool-use capabilities**.

Our core insight is that within a diverse pool of candidate proxy tasks, some tasks, although not directly involving tool use, exhibit structural or cognitive resemblance. Performance on these tasks during early training can provide predictive signals of eventual tool-use capability. To systematically identify effective proxy tasks, we introduce two key metrics: *alignment*, which measures the correlation between task performance and final tool-use ability across models and configurations; and *consistency*, which quantifies the stability of the correlation across training process. We develop concrete procedures for computing these metrics, enabling principled and scalable selection of proxy tasks.

Leveraging these two metrics as task-level weights, we develop a framework that aggregates early-stage proxy-task performance to predict each model's eventual tool-use capability. Given a collection of small- to medium-scale LLMs trained under a fixed setup, we compute weighted proxy-task scores at an early checkpoint and use them to infer the models' final relative rankings based on tool-use performance. Across diverse benchmarks, model scales, data mixtures, and hyperparameter settings, our approach consistently produces robust and significantly more accurate rankings compared to standard baselines such as early loss and scaling extrapolation (Huang, Zhang, and Liu 2024; Maddison, Ruan, and Hashimoto 2025). These results demonstrate our framework's effectiveness in anticipating emergent tool-use capabilities well before training convergence.

By exposing the limitations of existing forecasting methods, this study presents a theoretically grounded and practically effective framework for early prediction. Our findings facilitate more efficient model development and capability planning, while simultaneously advancing the theoretical understanding of emergent behaviors in LLMs.

*Corresponding Author.

Problem Formulation

We consider the problem of predicting the relative ordering of a collection of LLMs according to tool-use capabilities at the final training stage T, based solely on early-stage performances at an intermediate checkpoint t_0 on a collection of non-emergent proxy tasks $\mathcal{P} = \{p_1, p_2, \ldots, p_k\}$.

Setup. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ denote a set of LLM training configurations, where each c_i corresponds to a specific combination of model- or optimization-related parameters, such as model size, data mixture, hyperparameters, etc.

Here, we define the following functions:

- $f_T : \mathcal{C} \to R$, the *target capability function* mapping each configuration c_i to final tool-use performance $f_T(c_i)$ at the end of training step T. The objective is to predict the *ranking* induced by f_T, i.e., the ordering of $\{c_i\}$ based on ultimate tool-use abilities, without observing f_T.
- $g_{t_0}^{(j)} : \mathcal{C} \to R$, the *performance function* mapping each configuration c_i to the performance of j-th proxy task measured at an early checkpoint t_0, where $j = 1, \ldots, k$.

The intermediate checkpoint $t_0 < T$ is designated after the initial burn-in phase, at which training enters a relatively stable regime: the hyperparameters, data mixture, and model architecture are **fixed and remain unchanged**. This ensures that proxy task performance observed at t_0 is meaningful and predictive for extrapolating capability trajectories.

Prediction Objective. Our goal is to learn a *predictor*

$$\hat{f}_T : R^k \to R$$

that maps the vector of proxy performances at t_0 into a scalar predictive score $\hat{f}_T(c_i)$, i.e., for each configuration c_i,

$$\hat{f}_T(c_i) = \hat{f}_T\big(g_{t_0}^{(1)}(c_i), g_{t_0}^{(2)}(c_i), \ldots, g_{t_0}^{(k)}(c_i)\big).$$

This score serves as an approximation to $f_T(c_i)$, in the terms of preserving their relative ordering across configurations.

Formally, we evaluate the quality of $\hat{f}_T$ through its **ranking consistency** with f_T: we require that for most pairs $(c_i, c_j), 1 \le i, j \le n$,

$$f_T(c_i) \preceq f_T(c_j) \iff \hat{f}_T(c_i) \preceq \hat{f}_T(c_j).$$

Core Assumptions. The following conditions are assumed: (1) training configurations—including hyperparameters, data distributions, and model architectures—remain **fixed and stable** between checkpoint t_0 and final step T; (2) model performance is **consistent and reproducible** across repeated runs under identical settings; (3) performance on **non-emergent proxy tasks** follows **smooth, log-linear scaling laws** enabling reliable extrapolation from t_0 to T; and (4) there exist **informative proxy tasks** whose early metrics **strongly correlate** with final tool-use ability, providing actionable predictive signals.

Key Challenges. Despite these assumptions, accurately predicting the final ranking of tool-use capabilities remains non-trivial. First, the latent and potentially complex relationship between proxy tasks and tool-use leads to **correlation identification** challenge. Second, **signal robustness** poses a significant barrier: certain proxy tasks may produce noisy or unstable measurements across varying training conditions. Lastly, ranking is highly **sensitive to small prediction deviations**, where minor estimation errors can disproportionately impact the fidelity.

Proxy-Guided Prediction Framework

In this section, we introduce a proxy-guided framework to predict emergent tool-use capabilities in LLMs. The key idea is to identify early proxy tasks whose training trajectories exhibit predictive signals for eventual capability. We develop a systematic pipeline that selects proxies based on alignment with the target ability and consistency across training variations. These signals are aggregated to yield final predictions. We present the framework, detail proxy selection principles, describe the prediction process, and provide theoretical insights supporting its validity.

Framework Overview

As illustrated in Figure 1, the proxy-guided prediction framework consists of three key stages designed to systematically anticipate emergent tool-use capabilities in LLMs.

Candidate Proxy Task Curation involves the careful compilation and organization of a diverse pool of candidate proxy tasks. We begin by defining a comprehensive taxonomy of LLM capabilities, including problem-solving, language understanding, knowledge retrieval, and reasoning. Leveraging this taxonomy, we collect and curate an extensive benchmark suite covering a wide spectrum of cognitive and linguistic skills. This curated candidate pool forms the foundation for subsequent analysis and serves as the basis for identifying proxy tasks that may provide predictive signals for emergent tool-use abilities.

Proxy Task Reliability Assessment focuses on quantifying the informativeness and robustness of each candidate proxy task. We introduce two complementary metrics: *alignment*, which measures the correlation between proxy task performance and final tool-use capability across a broad range of LLMs, and *consistency*, which evaluates the stability of proxy task signals across different training runs and setups. The details of these metrics and computation are described in the following section on proxy task selection.

Finally, **Emergent Tool-Use Prediction** leverages the weighted ensemble of selected proxy tasks to produce reliable predictions of a model's tool-use proficiency before the ability visibly emerges. This prediction stage integrates proxy task signals according to their assessed reliability, enabling accurate ranking of models under diverse training regimes. The methodology of this prediction process are presented in the subsequent section on tool-use prediction.

Together, these stages establish a principled and scalable framework for anticipating complex emergent behaviors in LLMs well ahead of conventional evaluation milestones.

Proxy Task Selection: Alignment and Consistency

Effective proxy-guided prediction requires selecting proxy tasks that offer reliable signals. We propose two metrics:

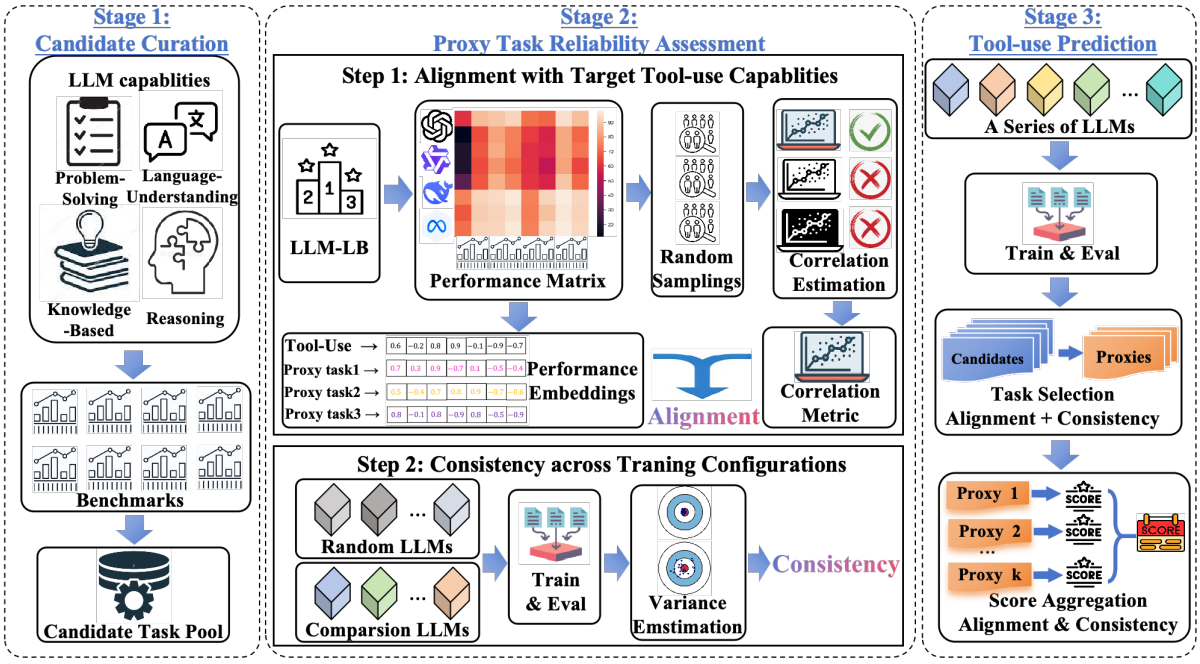

Figure 1: An overview of our proxy-guided framework for tool-use prediction, comprising three stages: (1) construction of proxy candidates from diverse LLM capability domains, (2) computation of predictive metrics, alignment (via correlation and performance embeddings) and consistency (via variance under training perturbations), and (3) aggregation of weighted proxy scores to estimate tool-use capability.

alignment, measuring how well proxy task performance correlates with final tool-use ability across configurations, and consistency, assessing the stability of this correlation across training checkpoints and setups. These criteria guide the selection of proxy tasks for accurate prediction.

Alignment Metric Computation. Accurately measuring the relevance between proxy tasks and the final tool-use capability is critical for effective proxy task selection. We utilize publicly available large-scale LLM leaderboard data to construct a diverse model set

$$\mathcal{M} = \{m_1, m_2, \ldots, m_m\},$$

and define the proxy task set as

$$\mathcal{P} = \{p_1, p_2, \ldots, p_k\}.$$

These leaderboards (e.g., HELM, Open LLM Leaderboard, Big-Bench) report evaluation scores of a wide range of large language models (LLMs) across diverse benchmark tasks, covering various capabilities such as reasoning, coding, and retrieval. Based on the scores reported in these leaderboards, we construct a *performance matrix* where each row corresponds to an LLM and each column corresponds to a task.

The performance vector across models of each proxy task p_j is

$$\mathbf{g}^{(j)} = \big(g^{(j)}(m_1), g^{(j)}(m_2), \ldots, g^{(j)}(m_m)\big) \in R^m,$$

and the target tool-use performance vector is

$$\mathbf{f} = \big(f_T(m_1), f_T(m_2), \ldots, f_T(m_m)\big) \in R^m.$$

To ensure comparability across tasks and models, we firstly standardize each proxy task vector to have zero mean and unit variance. Then for each model, the resulting values across tasks are further standardized. Let $\hat{\mathbf{g}}^{(j)}$ denote the doubly normalized performance vector for proxy task p_j, and let $\hat{\mathbf{f}}$ be the normalized target vector.

To quantify how well a proxy task aligns with the tool-use capability, we compute the correlation between its performance vector $g^{(j)}$ and the target vector f. We consider a set of candidate correlation functions

$$\mathcal{R} = \{\rho_1, \rho_2, \ldots, \rho_L\},$$

including commonly used statistical metrics such as **Pearson correlation**, **Spearman's rank correlation**, and **Kendall's tau**. The optimal alignment metric ρ^* is selected by evaluating its robustness across different settings.

Specifically, correlation metrics may behave differently depending on the number and identity of the models in the comparison set, and can be sensitive to sampling bias. Therefore, we propose two evaluation criteria to identify a robust and generalizable alignment metric:

Firstly, randomly sample S model subsets $\{\mathcal{M}_s\}_{s=1}^{S}$ of size $M_s < M$, and calculate:

$$r_l^{(s)} = \rho_l(\hat{\mathbf{g}}^{(j)}|_{\mathcal{M}_s}, \hat{\mathbf{f}}|_{\mathcal{M}_s}), \quad s = 1, \ldots, S.$$

Denote the correlation scores computed on the full model set $\mathcal{M}$, serving as the reference baseline and $\sigma(\mathbf{r}, t)$ as the set of top-t proxy tasks ranked by correlation scores in $\mathbf{r}$.

Baseline robustness measures average overlap between baseline and sampled top-t sets:

$$u_l = \frac{1}{S}\sum_{s=1}^{S}\frac{|\sigma(\mathbf{r}_l^{(0)},t)\cap\sigma(\mathbf{r}_l^{(s)},t)|}{t}.$$

Sampling robustness measures average overlap among sampled rankings:

$$v_l = \frac{2}{S(S-1)}\sum_{1\le s_1<s_2\le S}\frac{|\sigma(\mathbf{r}_l^{(s_1)},t)\cap\sigma(\mathbf{r}_l^{(s_2)},t)|}{t}.$$

The optimal correlation metric ρ^* is selected by maximizing both s_l and r_l:

$$u^* = \max_l u_l, \quad v^* = \max_l v_l.$$

The alignment score for proxy task p_j is then

$$\text{align}^{(j)} = \rho^*(\hat{\mathbf{g}}^{(j)}, \hat{\mathbf{f}}).$$

This method integrates comprehensive leaderboard data, rigorous normalization, and sampling-based validation to ensure a robust and principled alignment metric.

Consistency Metric Computation. In proxy task selection, besides evaluating the relevance between proxy tasks and the target task, assessing the robustness of proxy task performance under training uncertainties is equally critical. To this end, we propose a robustness analysis method based on two groups of small-scale models trained from scratch, enabling a principled quantification of task stability.

Specifically, we construct two small model ensembles: the *Data Variability Group* $\mathcal{D} = \{m_{d_1}, m_{d_2}, \ldots, m_{d_k}\}$ and the *Random Noise Group* $\mathcal{R} = \{m_{r_1}, m_{r_2}, \ldots, m_{r_k}\}$, where k denotes the number of models in each group. Both ensembles consist of small-scale models (e.g., ranging from 0.xB to under 10B parameters), which substantially reduces computational cost while retaining evaluation fidelity.

Within each group, training settings, except for the data source in $\mathcal{D}$ and the random initialization seeds in $\mathcal{R}$, are strictly aligned, including model architecture, training hyperparameters, and optimization schemes. In the Data Variability Group, models are trained on distinct datasets $\{D_1, D_2, \ldots, D_k\}$ to induce data distribution variability. Conversely, models in the Random Noise Group are trained on the same dataset but differ only in their random seeds $\{s_1, s_2, \ldots, s_k\}$, isolating stochastic training noise effects.

For each proxy task $p_j \in \mathcal{P}$, we collect performance vectors over two small model ensembles: the Data Variability Group $\{m_{d_1}, \ldots, m_{d_k}\}$ and the Random Noise Group $\{m_{r_1}, \ldots, m_{r_k}\}$, denoted as $\mathbf{g}_d^{(j)}$ and $\mathbf{g}_r^{(j)}$, respectively.

We then calculate the sample variance of the performance scores within each group. Let $\sigma_d^2(p_j)$ and $\sigma_r^2(p_j)$ denote the variances of $\mathbf{g}_d^{(j)}$ and $\mathbf{g}_r^{(j)}$, respectively.

The *consistency score* for proxy task p_j is defined as the ratio between these variances:

$$\text{consist}^{(j)} = \frac{\sigma_d^2(p_j)}{\sigma_r^2(p_j)}.$$

A higher consistency score indicates that the task's performance is more influenced by data distribution differences than by random initialization noise, implying greater robustness and reliability as a proxy task.

Moreover, this methodology can be generalized by replacing the Data Variability Group with ensembles varying in other controlled training factors (e.g., model architecture with fixed parameter sizes, or varying batch sizes) while maintaining data and random seed alignment. This enables multidimensional assessment of task robustness against diverse training perturbations.

This consistency metric effectively identifies proxy tasks whose evaluation results remain stable under inherent training uncertainties, providing dependable feedback for emergent ability prediction during early training phases.

Emergent Tool-Use Prediction

We integrate the alignment and the consistency metrics, forming a comprehensive metric for optimal proxy task selection and subsequent capability prediction.

Given a set of training configurations $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$, proxy task set $\mathcal{P} = \{p_1, p_2, \ldots, p_k\}$, and target task performance function $f_T : \mathcal{C} \to R$, we compute for each proxy task p_j two key scores: the alignment score $\text{align}^{(j)}$, quantifying the correlation between p_j and the target task f_T, and the consistency score $\text{consist}^{(j)}$, measuring the robustness of p_j.

To ensure the reliability of selected proxy tasks, we impose thresholds ϵ_C and ϵ_R on alignment and consistency scores respectively, retaining only tasks satisfying

$$\text{align}^{(j)} \ge \epsilon_C, \quad \text{consist}^{(j)} \ge \epsilon_R.$$

For each proxy task p_j passing the thresholds, we apply a monotonic transformation $f : R^+ \to [0,1]$ (e.g., sigmoid) to its consistency score to smooth the impact of extreme values. The final normalized weight is computed as

$$w^{(j)} = \frac{\text{align}^{(j)} \cdot f\big(\text{consist}^{(j)}\big)}{\sum_{p_i\in\mathcal{S}} \text{align}^{(i)} \cdot f\big(\text{consist}^{(i)}\big)}, \quad p_j \in \mathcal{S},$$

where $\mathcal{S}$ is the set of proxy tasks satisfying the thresholds.

The predicted tool-use capability for each training configuration c_i is given by the weighted sum of early-stage proxy task performances:

$$\hat{f}_T(c_i) = \sum_{p_j\in\mathcal{S}} w^{(j)} \cdot g_{t_0}^{(j)}(c_i), \quad \forall c_i \in \mathcal{C}.$$

Finally, the configurations $\{c_i\}$ are ranked by their predicted scores $\hat{f}_T(c_i)$, which serves as an approximation to the true ranking induced by $f_T(c_i)$. This ordering facilitates early-stage model selection by effectively predicting relative tool-use capabilities at the final training stage.

Theoretical Foundations

The feasibility of predicting tool-use capabilities from early-stage proxy task performance is theoretically plausible under reasonable conditions. Specifically, prior work has shown

that many non-emergent tasks exhibit smooth and monotonic learning curves that can be extrapolated from early checkpoints (Kaplan et al. 2020; Hernandez et al. 2021; Hoffmann et al. 2022).

Let $f_{t_0}^{(j)}$ and $f_T^{(j)}$ denote the performance of a model on proxy task j at early checkpoint t_0 and final step T, respectively. For non-emergent proxy tasks, prior work demonstrates that $f^{(j)}$ evolves smoothly under log-linear trends:

$$f_T^{(j)} \approx f_{t_0}^{(j)} + \beta_j \cdot \log(T/t_0),$$

for some task-specific slope β_j (Hernandez et al. 2021). This enables us to approximate final performance using early measurements.

Given that tool-use capabilities g_T may not be directly predictable from $f_{t_0}^{(j)}$, we posit that the relevant proxy tasks span a latent subspace $\mathcal{Z}$ shared with tool-use representations. Formally, let $g_T \in \mathcal{Z}$ and $\{f_T^{(j)}\}_{j=1}^m \subset \mathcal{Z}$. Then a linear combination of proxy scores can approximate g_T:

$$g_T \approx \sum_{j=1}^{m} w^{(j)} f_T^{(j)}.$$

Weights $w^{(j)}$ can be interpreted as reflecting each proxy task's *alignment* with and *consistency* across training configurations. Assuming a linear relation with additive noise,

$$f_{t_0}^{(j)}(c_i) = \alpha_j f_T(c_i) + \epsilon_i^{(j)},$$

where $\alpha_j > 0$ captures alignment and $\epsilon_i^{(j)}$ models noise with zero mean and variance inversely related to consistency. Under this model, the combined signal-to-noise ratio

$$\mathrm{SNR}_j = \frac{\sigma^2[\alpha_j f_T(c)]}{\sigma^2[\epsilon^{(j)}]}$$

justifies weighting proxy tasks to maximize predictive ranking fidelity. Empirically, this weighting preserves the relative ordering of configurations according to g_T, minimizing ranking errors (Fürnkranz and Hüllermeier 2010).

Results

Experimental Setup

We evaluate our proxy-based method for predicting tool-use capabilities across LLMs with different scales, datasets, and training setups. The experimental settings detail below.

Candidate Proxy Tasks. We select a suite of 42 benchmarks as proxy tasks spanning five capability categories: problem-solving (e.g., C-Eval, MMLU, GAOKAO-Bench), language (e.g., WiC, AFQMC, Flores), knowledge (e.g., TriviaQA, NaturalQuestions), comprehension (e.g., RACE, CSL, LAMBADA), and reasoning (e.g., ReCoRD, GSM8K, BBH, HumanEval).

Benchmarks and Evaluation. The pretrained models are fine-tuned with LoRA (Hu et al. 2021) on the Glaive function call dataset (GlaiveAI 2023) for 5 epochs. Evaluation is conducted on two benchmarks: T-Eval (Chen et al. 2023b), which measures under a ReAct framework (Yao et al. 2023) limited to 20 dialogue turns, and CIBench (Zhang et al. 2024a), which assesses end-to-end and oracle task execution.

Alignment Computation. We construct a performance matrix from OpenCompass (Contributors 2023) and the Open LLM Leaderboard (Fourrier et al. 2024), covering 34 models across 17 base and instruction-tuned variants. These include representative families such as LLaMA-2 (Touvron et al. 2023), Qwen (Bai et al. 2023), etc. An excerpt of the resulting matrix is shown in Table 1, where each row corresponds to a model and each column to a benchmark task. To identify the most predictive proxy tasks and the optimal correlation metric, we compare Pearson, Spearman, and Kendall's tau across various sampling configurations, varying task count ($S \in \{15, 25, 35\}$), model subset size ($M_s \in \{6, 8, 10\}$), and top $t = 10$ scoring focus.

Model	HumanEval	BBH	GSM8K
LLaMA2-13B	29.1	34.7	48.2
Qwen-7B	35.3	40.0	55.9

Table 1: Excerpt of the model-task performance matrix.

Consistency Computation. The two small model groups share Qwen-1.8B architectures, hyperparameters, and optimization settings to ensure consistent training conditions (Bai et al. 2023). The **Data Variability Group** comprises five models, each trained on distinct 100B-token subsets drawn from public datasets including Falcon RefineWeb (Research 2023), RedPajama-v2 (Computer 2023), Wikipedia (Foundation 2023), BookCorpus (Zhu et al. 2015), and CodeParrot (Face 2022), introducing controlled data distribution variation. The **Random Noise Group** consists of three models trained on different 30B-token subsets of Falcon RefineWeb, each derived from unique random shuffles to isolate stochastic training noise effects.

Tool-use Prediction. Tool-use prediction experiments use LLMs with 1.8B and 7B parameters, each trained on 3.6TB tokens under consistent model architecture. The 1.8B model is configured with 24 layers, 2048 hidden dimension, 2048 context length, and 5504 intermediate dimension; the 7B model uses 32 layers, 4096 hidden dimension, 4096 context length, and 14336 intermediate dimension. Both models use QKV bias, 32 attention heads, 32 KV groups, a peak learning rate of 1.2e-3, and a batch size of 12M tokens. We evaluate 6 training configurations derived from 2 learning rate schedules and 3 data curation pipelines. The schedules include a standard cosine decay (Loshchilov and Hutter 2016) and a cosine decay with final-stage linear annealing. All models are trained on subsets of a 4.5TB-token corpus covering web documents, code, books, encyclopedias, QA, and academic papers. The data curation pipelines are as follows: (1) Perplexity filtering, which retains low-perplexity domains using a reference language model (3.6TB); (2) Multi-criteria filtering, which applies content safety checks and education-level constraints; and (3) Diversified filtering, which extends (2) with explicit source balancing and domain diversification. Proxy-task performance is evaluated at four checkpoints (0.5T, 2.4T, 3.2T, and 3.6T tokens) to differentiate training configurations, with tool-use abilities remaining negligible before fine-tuning.

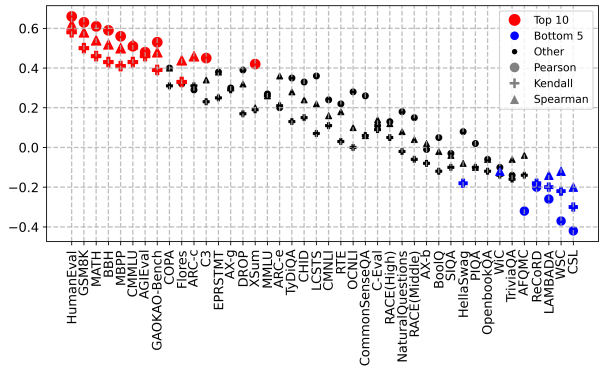

Figure 2: Alignment score between proxy tasks and T-eval benchmark. Markers indicate correlation types (Pearson, Spearman, Kendall), while color and size represent task ranking groups: top 10 (large, red), bottom 5 (medium, blue), and intermediate rankings (small, gray). Correlation values range from –0.5 to +0.7.

Hardware Infrastructure. Evaluation for tool-use benchmarks are conducted on a server with 8 NVIDIA A100 GPUs. Other pretraining experiments are executed on a GPU cluster composed of 60*8 NVIDIA H100 GPUs.

Main Results

Alignment Results and Insights. We compute alignment between 42 candidate proxy tasks and the T-eval tool-use benchmark using Pearson, Spearman, and Kendall correlation coefficients, with results summarized in Figure 2. Consistently across all three correlation metrics, the top-10 proxy tasks are predominantly reasoning tasks (including HumanEval, GSM8k, BBH, MATH, and MBPP), and problem-solving tasks such as CMMLU, AGIEval, GAOKAO-Bench, and ARC-c. Comprehension tasks like C3 and XSum also contribute to the upper ranks, albeit less prominently. Language tasks generally appear less predictive, with exceptions like Flores occasionally reaching higher ranks. Tasks consistently ranking in the bottom five include comprehension task CSL and language tasks WSC and LAMBADA. This distribution aligns closely with established task categorizations, underscoring that robust reasoning and problem-solving capabilities are the strongest predictors of tool-use performance in LLMs. The prominence of these task categories reflects the intrinsic complexity and multi-step reasoning demands of tool-use benchmarks like T-eval, as well as the structural similarity between tool-use and problem-solving tasks. This evidence substantiates the selection of reasoning and problem-solving tasks as reliable proxy indicators in early-stage emergent ability prediction.

Robustness Evaluation of Correlation Metrics. The robustness of proxy-task alignment metrics are analyzed under two settings: baseline robustness and sampling robustness. We vary the number of models M_s, number of sampled proxy tasks S, and report Pearson, Spearman, and Kendall correlations with final tool-use rankings. The results are shown in Table 2. We observe three notable trends across

(S, M_s)	BR-P	BR-S	BR-K	SR-P	SR-S	SR-K
(6, 25)	0.444	0.444	**0.492**	0.359	0.325	**0.372**
(8, 25)	0.544	0.516	**0.548**	0.418	0.392	**0.431**
(10, 25)	0.500	0.568	**0.580**	**0.476**	0.472	0.475
(10, 15)	0.560	0.613	**0.640**	0.467	0.491	**0.522**
(10, 25)	0.500	0.568	**0.580**	**0.476**	0.472	0.475
(10, 35)	0.551	**0.574**	**0.574**	0.435	0.434	**0.457**

Table 2: Robustness evaluation of correlation metrics under different experimental configurations. BR = Baseline Robustness, SR = Sampling Robustness, P=Pearson, S=Spearman, K=Kendall, top-t=10 scores are focused.

robustness evaluations. First, Kendall's rank correlation consistently yields more stable rankings than Pearson or Spearman, especially under low sampling settings. This reflects its robustness to noise and its suitability for capturing general monotonic trends between proxy and target tasks. Second, increasing the number of models M_s improves correlation in both baseline and sampling conditions, indicating that a broader evaluation set enhances ranking stability. Third, smaller but carefully chosen proxy task sets (e.g., S=15) often outperform larger ones (e.g., S=35), suggesting that excessive diversity may introduce noise. Overall, Kendall's metric aligns well with the inherent nonlinearity and distributional heterogeneity of language tasks, making it well-suited for alignment computation.

Task	RN	DV	Consistency	Highlight?
C3	0.91	51.97	**57.23**	Y
CMNLI	0.40	1.18	2.94	N
OCNLI	8.23	7.20	0.88	N
CHID	11.27	577.55	**51.23**	Y
RTE	1.71	8.37	4.90	-
CMMLU	0.04	0.39	10.79	-

Table 3: Task performances under two small-scale model groups, including variances and consistency scores. RN=Random Noise Group, DV=Data Variability Group.

Consistency Results and Insights. The consistency of proxy tasks are based on early-stage LLM performance variance under two perturbations: Random Noise and Data Variability groups. Six tasks, C3, CMNLI, OCNLI, CHID, RTE, and CMMLU, were chosen for their high alignment with T-eval and stable early performance (unlike HumanEval, GSM8k, etc.), making them effective proxy candidates. These tasks cover diverse abilities including inference, reading comprehension, and QA, employing a multiple-choice format enabling perplexity-based evaluation. Results (Table 3) show reading comprehension tasks C3 and CHID achieve the highest robustness, indicating resilience to training noise. CMMLU's consistent difficulty yields low variance but high robustness. Among inference tasks, RTE demonstrates moderate stability, whereas CMNLI and OCNLI exhibit lower robustness due to higher sensitivity to training perturbations. These findings validate robustness estimation as a practical complement to correlation metrics for selecting reliable proxy tasks in early stages.

1.8B	C+P	C+M	C+D	A+P	A+M	A+D	R-T	R-C
0.5T	45.01	45.40	**45.57**	45.01	45.40	**45.57**	5/15	5/15
2.4T	45.95	46.40	**47.20**	45.95	46.40	**47.20**	5/15	5/15
3.2T	46.60	47.10	**47.77**	47.25	47.76	**48.03**	1/15	3/15
3.6T	46.52	46.75	**47.69**	46.80	47.72	**47.91**	**0/15**	**2/15**
T-eval	21.90	22.30	**22.76**	22.35	23.20	**23.86**	0/15	-
CIBench	17.42	18.05	**18.30**	17.80	**18.62**	18.60	-	0/15
7B	C+P	C+M	C+D	A+P	A+M	A+D	R-T	R-C
0.5T	47.22	**48.05**	47.75	47.22	**48.05**	47.75	3/15	3/15
2.4T	49.80	**50.92**	50.45	49.80	**50.92**	50.45	3/15	3/15
3.2T	50.83	**52.08**	51.78	51.15	**52.33**	51.60	**1/15**	**1/15**
3.6T	50.92	**52.12**	51.83	51.20	**52.38**	51.66	**1/15**	**1/15**
T-eval	56.20	**58.90**	58.63	57.24	**59.52**	58.80	0/15	-
CIBench	44.92	**48.00**	46.53	45.36	**47.14**	45.65	-	0/15

Table 4: Tool-use prediction results across 1.8B and 7B model groups, C=Cosine decay, A=C+final-stage linear annealing, P=Perplexity filtering, M=Multi-criteria filtering, D=Diversified filtering, R-T / R-C: Number of reversed pairs between predicted and ground truth (lower is better).

	1.8B R-T	1.8B R-C	7B R-T	7B R-C
Our method	**0/15**	**2/15**	**1/15**	**1/15**
w/o align	5/15	8/15	4/15	6/15
w/o consist	4/15	5/15	3/15	5/15
PPL	6/15	10/15	8/15	9/15
SL.	10/15	8/15	10/15	11/15
Avg.	6/15	8/15	5/15	5/15

Table 5: Ablation and comparison with baselines. PPL: prediction by validation perplexity; SL: scaling law extrapolation; Avg: average proxy task scores; w/o align / consist: without alignment / consistency metrics.

Tool-use Prediction Results. Table 4 presents the prediction accuracy of proxy-task-based ranking under six training configurations. For each benchmark, we report the number of reversed pairwise orderings out of 15 possible comparisons. Results show strong alignment between proxy-task predictions and final tool-use performance on both T-eval and CIBench. At the 3.6T checkpoint, the 1.8B model achieves 0 and 2 reversed pairs on T-eval and CIBench, while the 7B model obtains 1 reversed pair on both. The correct ordering of data curation strategies (P, M, D for 1.8B and 7B model) is consistently recovered, and the improvement of A over C is also captured at later stages. These results indicate that proxy tasks can reliably reflect global structure in model capability progression. Notably, proxy-task supervision generalizes across benchmarks. Although proxy weights are derived from T-eval alignment, the induced predictions remain accurate on CIBench. This suggests that well-chosen proxy tasks can encode transferable signals relevant to tool-use emergence. Early checkpoints provide partial predictability. While differences between optimization schedules are unclear at 0.5T and 2.4T, data-related variation is already captured. This implies that signal emergence in proxy tasks is axis-dependent, with data factors becoming predictable earlier than optimization choices.

Ablation and Comparison Results. As shown in Table 5, our method consistently achieves the lowest reversed pair counts across all settings, demonstrating superior ranking alignment. Removing either the alignment or consistency components degrades performance, confirming their complementary roles. Baselines relying solely on perplexity or scaling law extrapolation yield substantially higher errors, highlighting the advantage of our approach.

Related Work

Scaling Laws. Recent studies have revealed that the performance of LLMs on many tasks follows predictable scaling laws with respect to compute, model size, and data (Kaplan et al. 2020; Hoffmann et al. 2022; Hernandez et al. 2021). These trends often manifest as smooth, log-linear trajectories, particularly for non-emergent capabilities. Such observations form the foundation for extrapolating future performance from early checkpoints (Zhang et al. 2022).

Emergent Abilities in LLMs. A growing body of work has investigated the sudden appearance of complex behaviors, referred as emergent abilities, in LLMs as a function of scale (Wei 2022; Ganguli, Hernandez et al. 2022). These abilities often defy extrapolation from small models or early training stages, posing challenges for model evaluation and development. Tool-use is widely considered one such emergent capability, requiring models to interface with external APIs or environments for reasoning and execution (Schick et al. 2023; Liu et al. 2023).

Tool-Use Benchmarking. A number of recent benchmarks aim to systematically evaluate LLMs' ability to interact with tools, environments, or APIs (Qin et al. 2023; Parisi et al. 2022; Chen et al. 2023a). These studies typically measure success rates on planning, tool selection, and execution tasks. While some works explore fine-tuning or in-context adaptation for tool-use (Patil et al. 2023; Yao et al. 2022), few provide systematic methods to predict final tool-use capabilities from early-stage signals.

Proxy Tasks as Predictive Signals. Several efforts have explored using auxiliary tasks to guide or predict downstream performance (Abnar et al. 2022; Vu et al. 2020; Zamfirescu-Pereira et al. 2022). Prior work rarely identifies early-stage proxy tasks predictive of emergent tool-use in large models. Insights from other domains, including multi-scale feature fusion (Zhou et al. 2025), dynamic sampling (Zhang et al. 2024b), and cross-modal feature alignment (Fang et al. 2025), highlight the value of intermediate representations for designing predictive proxy tasks.

Conclusion

We present a principled framework to predict LLM tool-use emergence via early performance on non-emergent proxy tasks. Combining alignment-based selection with consistency-aware aggregation, our method is theoretically grounded and empirically validated. While effective, it assumes stable training dynamics and may need adaptation across models. Our approach provides a foundation for optimizing training decisions before emergent capabilities arise.

Acknowledgements

The research is supported National Science Fund for Distinguished Young Scholars(62125601), State Key Laboratory of Multimedia Information Processing Open Fund(SKLMIP-KF-2025-03).

References

Abnar, S.; et al. 2022. Quantifying memorization across training runs. *arXiv preprint arXiv:2202.07646*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Chen, J.; et al. 2023a. Tool-Enhanced Encoder-Decoder Transformers for LLM Tool-Use. *arXiv preprint arXiv:2310.02271*.

Chen, Z.; et al. 2023b. T-Eval: Evaluating the Tool Utilization Capability Step by Step. *arXiv preprint arXiv:2312.14033*.

Computer, T. 2023. RedPajama: an Open Dataset for Training Large Language Models. *arXiv preprint arXiv:2411.12372*.

Contributors, O. 2023. OpenCompass: A Universal Evaluation Platform for Foundation Models. https://github.com/open-compass/opencompass.

Face, H. 2022. CodeParrot Dataset.

Fang, Z.; Zhu, X.; Yang, C.; and et al. 2025. Aligning enhanced feature representation for generalized zero-shot learning. *Science China Information Sciences*, 68: 122102.

Foundation, W. 2023. Wikipedia Dataset on Hugging Face: Structured Content for AI/ML. *Hugging Face Datasets*.

Fourrier, C.; Habib, N.; Lozovskaya, A.; Szafer, K.; and Wolf, T. 2024. Open LLM Leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Fürnkranz, J.; and Hüllermeier, E. 2010. Preference Learning. In *Encyclopedia of Machine Learning*, 760–766. Springer.

Ganguli, D.; Hernandez, D.; et al. 2022. Predictability and surprise in large generative models. *arXiv preprint arXiv:2202.07785*.

GlaiveAI. 2023. Glaive Function Calling V2 Dataset. https://huggingface.co/datasets/glaiveai/glaive-function-calling-v2. Accessed: 2025-07-31.

Hernandez, D. J.; et al. 2021. Scaling Laws for Transfer. *arXiv preprint arXiv:2102.01293*.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; et al. 2022. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556*.

Hu, E. J.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Huang, Y.; Zhang, M.; and Liu, W. 2024. Predicting Downstream Performance in LLMs via Proxy Metrics and Scaling Laws. *Transactions of Computational Linguistics*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

Li, M.; Zhao, Y.; Yu, B.; Song, F.; Li, H.; Yu, H.; Li, Z.; Huang, F.; and Li, Y. 2023. API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs. In *Proceedings of EMNLP 2023*.

Liu, N. F.; et al. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Loshchilov, I.; and Hutter, F. 2016. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv preprint arXiv:1608.03983*.

Luo, N.; Gema, A. P.; He, X.; van Krieken, E.; Lesci, P.; and Minervini, P. 2025. Self-Training Large Language Models for Tool-Use Without Demonstrations. *arXiv preprint arXiv:2502.05867*.

Maddison, C.; Ruan, Y.; and Hashimoto, T. 2025. Scaling Laws for Predicting Downstream Performance in Large Language Models. In *ICLR 2025*.

Parisi, L.; et al. 2022. TOTA: Tool-Augmented Reasoning with Large Language Models. In *Proceedings of the NeurIPS*.

Patil, A.; et al. 2023. Gorilla: Large Language Model Connected with Massive APIs. *arXiv preprint arXiv:2305.15334*.

Qin, C.; et al. 2023. ToolBench: Towards unified benchmark for foundation model's tool-use. *arXiv preprint arXiv:2307.15602*.

Research, T. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only. *arXiv preprint arXiv:2306.01116*.

Schick, T.; Dwivedi-Yu, A.; Schütze, H.; et al. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Tang, Q.; Deng, Z.; Lin, H.; Han, X.; Liang, Q.; Cao, B.; and Sun, L. 2023. ToolAlpaca: Generalized Tool Learning for Language Models with 3000 Simulated Cases. *arXiv preprint arXiv:2306.05301*.

Touvron, H.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vu, X.-S.; et al. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the EMNLP*.

Wei, J. e. a. 2022. Emergent Abilities of Large Language Models. *arXiv preprint arXiv:2206.07682*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yao, S.; et al. 2022. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Zamfirescu-Pereira, M.; et al. 2022. Taskonomy of language model behaviors. *arXiv preprint arXiv:2209.00046*.

Zhang, C.; et al. 2024a. CIBench: Evaluating the Code Interpreter Capability of Large Language Models Step by Step. *arXiv preprint arXiv:2407.10499*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; Mihaylov, T.; Ott, M.; Shleifer, S.; Shuster, K.; Simig, D.; Koura, P. S.; Sridhar, A.; Wang, T.; and Zettlemoyer, L. 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*.

Zhang, S.-X.; Yang, C.; Zhu, X.; Zhou, H.; Wang, H.; and Yin, X.-C. 2024b. Inverse-Like Antagonistic Scene Text Spotting via Reading-Order Estimation and Dynamic Sampling. *Trans. Img. Proc.*, 33: 825–839.

Zhou, H.; Zhu, X.; Qin, J.; Xu, Y.; Cesar-Jr, R. M.; and Yin, X.-C. 2025. Multi-Scale Texture Fusion for Reference-Based Image Super-Resolution: New Dataset and Solution. *International Journal of Computer Vision*, 133: 6971 – 6992.

Zhu, Y.; Lin, C. P.; Bhat, M. R. G.; and S., R. S. S. S. S. H. J. 2015. BookCorpus Dataset.