How Much Do Large Language Model Cheat On Evaluation? Benchmarking Overestimation Under The One-Time-Pad-Based Framework

Zi Liang[1], **Liantong Yu**[1], **Shiyu Zhang**[1], **Qingqing Ye**[1], **Haibo Hu**[1,2*]
[1]The Hong Kong Polytechnic University
[2]PolyU Research Centre for Privacy and Security Technologies in Future Smart Systems
{zi1415926.liang,liantong2001.yu,shiyu187.zhang}@connect.polyu.hk,
{qqing.ye,haibo.hu}@polyu.edu.hk

Abstract

Overestimation in evaluating large language models (LLMs) has become an increasing concern. Due to the contamination of public benchmarks or imbalanced model training, LLMs may achieve unreal evaluation results on public benchmarks, either intentionally or unintentionally, which leads to unfair comparisons among LLMs and undermines their realistic capability assessments. Existing benchmarks attempt to address these issues by keeping test cases permanently secret, mitigating contamination through human evaluation, or repeatedly collecting and constructing new samples. However, these approaches fail to ensure reproducibility, transparency, and high efficiency simultaneously. Moreover, the extent of overestimation in current LLMs remains unquantified. To address these issues, we propose ArxivRoll, a dynamic evaluation framework inspired by one-time pad encryption in cryptography. ArxivRoll comprises two key components: *i) SCP (Sequencing, Cloze, and Prediction)*, an automated generator for private test cases, and *ii) Rugged Scores (RS)*, metrics that measure the proportion of public benchmark contamination and training bias. Leveraging SCP, ArxivRoll constructs a new benchmark every six months using recent articles from ArXiv and employs them for one-time evaluations of LLM performance. Extensive experiments demonstrate the high quality of our benchmark, and we provide a systematic evaluation of current LLMs.

1 Introduction

With the rapid development of large language models (LLMs), their evaluation has attracted growing attention. Numerous challenging and widely recognized benchmarks (Hendrycks et al. 2021; Cobbe et al. 2021; White et al. 2024; Chiang et al. 2024; Jimenez et al. 2024; Team 2025) have been introduced to assess the knowledge and reasoning capabilities of these models. As a result, these evaluations have become the primary, and often the only, standard for comparing the performance of large language models.

Despite their effectiveness, recent research (Wu et al. 2024; Dong et al. 2024; Jiang et al. 2024) increasingly highlights the shortcomings of current evaluation mechanisms, arguing that the capabilities of LLMs are often universally *overestimated*. This occurs mainly due to evaluation leakage, where test samples, benchmark details or formatting information can be exploited to game the benchmark. Consequently, it may inflate the perceived performance of a model, resulting in unreliable evaluations and unfair comparisons among LLMs. Malicious developers could further fool benchmarks by incorporating test samples or benchmark-specific information during training or fine-tuning. For instance, a previous study (Yang et al. 2023) demonstrated that a 13-billion-parameter Llama model can easily achieve results comparable to GPT-4 on benchmarks like MMLU (Hendrycks et al. 2021) through post-processing-based fine-tuning. Additionally, popular open-source LLMs such as Llama-4 and Qwen-2.5 have been reported (Tech Startups 2025; Wu et al. 2025) to experience test-data-contaminated training. Such intentional or unintentional cheating behaviors distort the true capabilities of LLMs, misleading subsequent training procedures and corresponding discoveries (Wu et al. 2025).

Specifically, there are two main types of abuse involving evaluation benchmarks. The first is *data contamination* (Palavalli, Bertsch, and Gormley 2024; Li et al. 2024c; Dong et al. 2024; Xu et al. 2024; Jiang et al. 2024), where test cases from the benchmarks are included in the training set of large language models, enabling them to become familiar with or even memorize these samples, resulting in artificially improved performance. The second is *biased overtraining*, where models are claimed to be "comprehensive" but actually prioritize improving their performance in the evaluated domain at the expense of undertraining in other areas. Both scenarios significantly undermine the effectiveness, fairness, and reliability of evaluation results.

Unfortunately, existing benchmarks designed to mitigate cheating behaviors have notable limitations. Private benchmarks maintained by trusted third-party platforms, such as SEAL[1], and Arena-like benchmarks (Chiang et al. 2024; Huang et al. 2024; Li et al. 2024b,a), such as Chatbot Arena (Chiang et al. 2024), lack transparency and reproducibility in their evaluation processes. Symbolic formatting benchmarks for specific domains (Zhu et al. 2024a; Zhang, Chen, and Yang 2024; Zhu et al. 2024b), such as GSM-Symbolic (Mirzadeh et al. 2024) and LiveBench (White et al. 2024), are restricted to narrow fields and therefore fail to provide a comprehensive evaluation of LLMs. Furthermore,

*Corresponding author.

[1]https://scale.com/leaderboard

the above benchmarks primarily focus on assessing the realistic abilities of LLMs, without offering a clear *quantification* of the extent of overestimation. As a result, a stable, transparent, reproducible, and human-effort-free framework and benchmark for evaluating LLMs has yet to be developed.

To address these issues, we propose ArxivRoll, a robust and dynamic framework designed to evaluate both the realistic performance and the overestimation of large language models. ArxivRoll consists of two key components: *1)* SCP (Sequencing, Cloze, and Prediction), a novel method that automatically generates test cases from newly published articles on ArXiv to construct private benchmarks; and *2)* Rugged Scores (RS), indicators that quantify the performance difference between public and private benchmarks, providing a clear measure of overestimation. Inspired by the security guarantee of *One-Time Pad* (Miller 1882; Shannon 1949) in cryptography, which uses a unique secret key for each use, ArxivRoll divides benchmarks into public benchmarks (existing ones) and private ArxivRollBenches (generated by SCP), and regard the private benchmarks as the **one-time-used** secrets to mitigate the overestimation. After evaluation, the private benchmarks are publicly released to ensure reproducibility of evaluation but are marked as expired to prevent future use or reference. Extensive meta-evaluations on ArxivRollBench demonstrate that SCP consistently produces high-quality test samples. Besides, the private benchmarks exhibit a strong correlation with existing private yet non-transparent benchmarks, confirming their reliability and relevance.

Our contributions are summarized as follows:

- We devise a novel private benchmark construction strategy, SCP (Sequencing, Cloze, and Prediction) based on Arxiv, which automatically generates high-quality, challenging, and fresh test cases tailored for assessing the capabilities of LLMs. Extensive experiments have proved the high quality of our generated private benchmarks.
- We design *rugged scores (RS)* to quantify the proportion of cheating behavior in a given LLM when tasked with specific challenges. To the best of our knowledge, this is the first study to measure the proportion of overestimation and the biased overtraining.
- Leveraging RS and SCP, we present a novel and one-time-pad-based evaluation framework, namely ArxivRoll. This framework not only evaluates the performance of current LLMs but also considers their overestimation situations. Through ArxivRoll, we conduct a systematic evaluation and establish a leaderboard[2] for popular LLMs, providing a comprehensive evaluation of their capabilities.

The source code is available at https://github.com/liangzid/ArxivRoll/.

2 ArxivRoll

In this section, we will introduce the implementation of ArxivRoll and explain why it can address limitations that existed in previous benchmarks. Specifically, we first introduce our dynamic evaluation framework in Section 2.1, and then respectively detail our test cases generator technique as well as the metrics in Section 2.2 and 2.3.

2.1 Overview

As illustrated in Figure 1, ArxivRoll encompasses two categories of benchmarks: *public* and *private*. Public benchmarks refer to those publicly available on the Internet, which may be susceptible to contamination or hacking during the pretraining of LLMs. Conversely, private benchmarks, namely ArxivRollBench, are created by ArxivRoll and remain confidential until the evaluation period, thereby ensuring that they are unseen by LLMs. In addition to assessing performance, ArxivRoll also computes two key values:

- The difference in performance for an LLM between the public and private benchmarks within the same domain (e.g., mathematics reasoning). This metric reflects the proportion of contamination in the model's performance on public benchmarks;
- The difference in performance for an LLM among various private benchmarks. This metric indicates the degree of biased overtraining in the model.

We propose the rugged score to quantify these two differences, as shown in Section 2.3.

After evaluation, we can compile the performances and rugged scores for all LLMs into a leaderboard and make our constructed private benchmarks publicly available on the Internet to ensure the reproducibility and transparency of the evaluation process. These benchmarks will be regarded as public benchmarks in future evaluations.

This outlines the entire procedure of ArxivRoll for one evaluation period. As a dynamic benchmark, it will regularly publish new evaluations (e.g., every six months). For each evaluation period, as shown above, ArxivRoll will incorporate new private benchmarks to minimize the impact of contamination and biased overtraining, as shown in Figure 1.

Such a framework faces two primary challenges:

- How do we create confidential benchmarks for each evaluation stage that are both challenging and representative of the domain, while ensuring they remain unseen by LLMs until the evaluation period?
- How do we formally measure the two differences to provide a rigorous and interpretable evaluation?

We will address them in the following two parts.

2.2 Sequencing, Cloze, and Prediction (SCP): Producing Test Cases

From Section 2.1, we can discern that private benchmarks must meet the following four criteria: *i) confidentiality*, ensuring that LLMs do not encounter the test cases during their training process; *ii) difficulty*, where test cases should not adhere to fixed patterns but remain flexible and complex in content, preventing LLMs from easily solving them through lexical comprehension alone; *iii) objectivity*, to minimize the impact of subjective evaluation metrics; and *iv) comprehensiveness*, for encompassing a wide range of fields or

[2]https://arxivroll.moreoverai.com

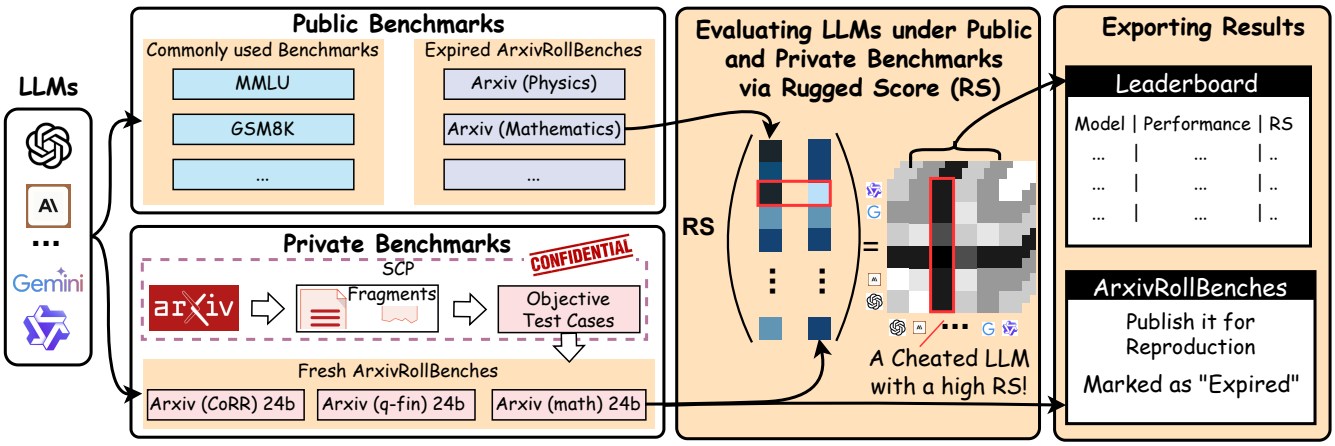


Figure 1: **Framework of ArxivRoll**, which categorizes benchmarks into two distinct groups: public benchmarks and private benchmarks (i.e., ArxivRollBench). These benchmarks are utilized to estimate both the overestimation proportion and performance of large language models (LLMs). Notably, ArxivRoll represents a *dynamic* benchmarking system, where private benchmarks are utilized exclusively once and then expire for subsequent evaluation stages, ensuring freshness and reliability in each assessment.

sub-fields rather than being confined to a narrow task. Moreover, given the need to introduce new private benchmarks for each evaluation stage, we aspire to construct test cases *automatically*.

To this end, we have chosen ArXiv[3], a preprint platform, as the source for our test cases. The timely papers published on ArXiv fulfill the criteria of confidentiality and difficulty, as they represent the latest research advancements in their domains and are often unprecedented in academia. Consequently, these papers are conceptually unseen by LLMs to date, making them suitable for our benchmark construction.

Despite the potential of designing test samples based on ArXiv articles, the process remains time-consuming and challenging, necessitating expert-level annotators. To tackle this issue, we adopt the concept of symbolic formatting and propose an automated test sample generation strategy named SCP. SCP is inspired by educational quizzes (Abraham and Chapelle 1992; Bormuth 1968; Alderson 1979) and Gestalt psychology (Britannica 2024; Mather 2006), which comprises three objective tasks:

- **Sequencing**: Given a text fragment extracted from an article, the input of a test case consists of shuffled sentences from this fragment. LLMs are tasked with selecting the correct order of these sentences.
- **Cloze**: In this task, a text fragment is provided with certain sentences masked. LLMs are required to select the appropriate sentences to fill in these gaps.
- **Prediction**: Given a text fragment, a correct subsequent sequence, and three distractors, LLMs must identify and select the correct next sequence.

Formally, for an article, we first sample a text fragment containing N_p paragraphs, filtering out texts heavy with mathematical formulas and tables. Then, we utilize one of the strategies within SCP to generate the test case. Figure 2 depicts the construction process of SCP.

[3]https://arxiv.org/

2.3 RS: Quantifying Overestimation

Given both public and private benchmarks, another challenge arises in assessing the reliability of performance evaluations conducted on public benchmarks. Intuitively, within the same domain, if an LLM demonstrates significantly higher performance on a public benchmark compared to a private one, we may conclude that the public benchmark is being "fooled" by the LLM. To quantify this discrepancy, we introduce a novel metric called the **rugged score (RS)**. This metric measures the degree of "ruggedness" in performance between public and private benchmarks.

Formally, given N_p public-private benchmark pairs $\mathcal{T} = \{(T_p^i, T_c^i)\}_{i=1,2,...,N_p}$ in which the public benchmark T_p^i and the private benchmark T_c^i comes from the same domain, and given N_p' unmatchable public benchmarks $\mathcal{T}_p = \{T_p^j\}_{j=1,2,...,N_p'}$ and N_c unmatchable private benchmarks $\mathcal{T}_c = \{T_c^k\}_{k=1,2,...,N_c}$, we can define the rugged score of a model m as:

$$\begin{aligned}&\mathbf{RS_I}(m, \mathcal{T}, \mathcal{T}_p, \mathcal{T}_c)\\ &= \frac{2}{N_p}\sum_i^{N_p}\left[\frac{\mathcal{M}(m, T_p^i) - \mathcal{M}(m, T_c^i)}{\mathcal{M}(m, T_p^i) + \mathcal{M}(m, T_c^i)}\right] + 2\times\\ &\left[\frac{\frac{1}{N_p'}\sum_j^{N_p'}\mathcal{M}(m, T_p^j) - \frac{1}{N_c}\sum_k^{N_c}\mathcal{M}(m, T_c^k)}{\frac{1}{N_p'}\sum_j^{N_p'}\mathcal{M}(m, T_p^j) + \frac{1}{N_c}\sum_k^{N_c}\mathcal{M}(m, T_c^k)}\right],\end{aligned} \tag{1}$$

where $\mathcal{M}(m, T)$ denotes the performance evaluation metric for the model m on task T. It can either be an absolute metric such as the accuracy, or a relative metric like the rank of m among all evaluated models M.

In intuition, the higher the $\mathbf{RS_I}$, the rugger the m, demonstrating that the evaluated results of m on public benchmarks

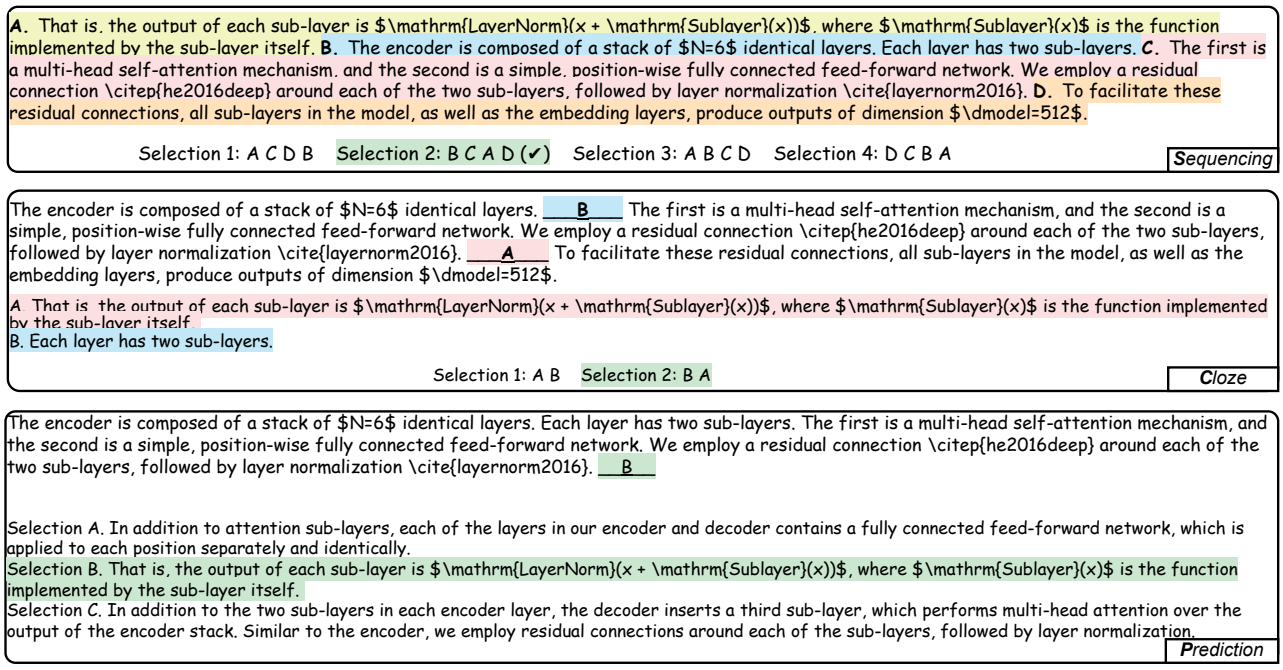


Figure 2: **An illustrative example of symbolic formatting for test samples**, encompassing three formats: sequencing, cloze, and prediction (SCP).

$\{T_p^i\}_{N_p} \cup \{T_p^j\}_{N_p'}$ may be less reliable, and model m may be overfitted to the specific characteristics of them.

Unfortunately, $\mathbf{RS_I}$ is not a *normalized* metric and is unavoidably coupled with models and benchmarks used for evaluation. This means that $\mathbf{RS_I}$ obtained for different sets of models M on different benchmark triples $(\mathcal{T}, \mathcal{T}_p, \mathcal{T}_c)$ are *incomparable*, and we can **only** decouple one factor between M and benchmarks triples from $\mathbf{RS_I}$. Specifically, $\mathbf{RS_I}$ becomes *model-independent* when an absolute metric is adopted as $\mathcal{M}$, allowing the free addition of new models under the same triple $(\mathcal{T}, \mathcal{T}_p, \mathcal{T}_c)$ without affecting the score's comparability. Conversely, it becomes *benchmark-independent* when a relative metric is used, meaning that it is comparable across different evaluation periods for the same model set M. In our evaluation, we will use both types of rugged scores.

To investigate the proportion of unbalanced overtraining on LLMs, we propose $\mathbf{RS_{II}}$, which can be measured by the standard variance on private benchmarks, i.e.,

$$\mathbf{RS_{II}} = \sqrt{\sum_{T_c \sim \{\mathcal{T}_c \cup \mathcal{T}_c^p\}} \left[\mathcal{M}(m, T_c) - \bar{\mathcal{M}}\right]^2}, \qquad (2)$$

where $\mathcal{T}_c^p = \{T_c^i, | i = 1, 2, ..., N_p\}$ represents the set of private benchmarks in $\mathcal{T}$, $\bar{\mathcal{M}} = \sum_{T_c \sim \{\mathcal{T}_c \cup \mathcal{T}_c^p\}} \mathcal{M}(m, T_c)$ is the average performance on private benchmarks. We also propose a normalized version:

$$\mathbf{RS_{II}^N} = \mathbf{RS_{II}} / \bar{\mathcal{M}}.$$

3 Meta Evaluation

In this section, we conduct a meta-evaluation of ArxivRoll. Specifically, Section 3.1 examines and assesses the quality of the generated test cases, while Section 3.2 investigates the potential correlation between ArxivRollBench's evaluation outcomes and those from other private benchmarks.

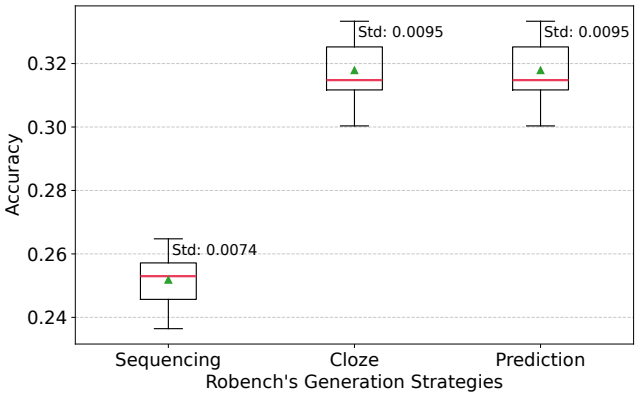


Figure 3: Performance of Llama3 (8B) across 32-time-generated ArxivRollBench benchmarks. The benchmarks were generated 32 times from the same raw article set using SCP. The small standard variance in evaluation results indicates that SCP produces stable test cases.

3.1 The Generation Of SCP Is Stable

Our benchmark construction strategy (SCP) relies heavily on randomness, raising whether the evaluation results adequately reflect an LLM's understanding of the articles. To address this, we examine if the performance of an LLM varies significantly when evaluated on the same set of raw articles but generated

Benchmarks	Spear. Corr.	Pearson Co.	Kendall Corr.
A.R.Bench (S) - ChatbotArena	0.76	0.71	0.6
A.R.Bench (C) - ChatbotArena	0.61	0.51	0.55
A.R.Bench (P) - ChatbotArena	0.73	0.69	0.55
A.R.Bench (S) - A.R.Bench (C)	0.86	0.92	0.77
A.R.Bench (S) - A.R.Bench (P)	0.86	0.86	0.69
A.R.Bench (C) - A.R.Bench (P)	0.86	0.86	0.69

Table 1: Correlation Experiments among ArxivRollBench (A.R.Bench) and ChatbotArena, where Spear. Corr., Pearson Co., and Kendall Corr. denote the Spearman Correlation, Pearson Coefficient, and Kendall Correlation, respectively.

with different random seeds.

Specifically, we repeated the generation process for ArxivRollBench2024b-CS 32 times using different seeds and collected the evaluation results of the Llama3-8B model. We calculated the performance variations across our three test case generation strategies: sequencing, cloze, and prediction, as illustrated in Figure 3.

Although the accuracy variations across all three benchmarks appear noticeable (~ 2.5 points), the standard deviation of the evaluation results is minimal, remaining below 1 point. This demonstrates that our generation strategy is reliable and that the evaluation results are consistently reproducible, even with benchmarks generated under varying random seeds.

3.2 ArxivRollBench Exhibits High Correlations With Popular Benchmarks

In this section, we address the second concern of ArxivRoll: whether the evaluation results from ArxivRollBench meaningfully reflect the knowledge and reasoning abilities of LLMs within these domains. To explore this, we compute the correlation between the performance rankings of ArxivRollBench2024b's private benchmarks and those of widely used benchmarks which are relatively harder to fool.

As a reference, we select ChatbotArena (Chiang et al. 2024), a crowdsourced and voting-based benchmark. Despite its limitations in interpretability, transparency, and reproducibility, ChatbotArena remains one of the most widely regarded benchmarks for LLM performance evaluation. We employ three standard correlation metrics: Pearson's coefficient, Spearman's rank correlation, and Kendall's rank correlation, all of which are commonly used to assess linear and rank-based relationships.

We first compute the correlations between our benchmark construction strategy, SCP, and the reference benchmark. Additionally, we analyze the internal correlations among the three test case generation strategies of SCP. The results of these analyses are presented in Table 1.

As shown in Table 1, ArxivRollBench constructed with the S (equencing) and P (rediction) strategies achieves up to a 0.70 Spearman correlation with ChatbotArena, while ArxivRollBench with the C (loze) strategy also exhibits a notable correlation with the reference benchmark. This demonstrates that ArxivRollBench's private benchmarks effectively capture the realistic capabilities of LLMs. Moreover, the strong correlations among the three SCP construction strategies indicate their internal consistency. The Pearson coefficients further suggest that their evaluation results exhibit linear relationships, reinforcing the robustness of our proposed approach.

Having established the utility of ArxivRollBench, we proceed to provide an in-depth analysis of the overestimation behavior of current LLMs under ArxivRoll in Section 4.

4 Evaluation

In this section, we evaluate current popular LLMs with ArxivRollBench and correspondingly quantify the proportions of the overestimation. Specifically, we first detail the settings of our evaluation in Section 4.1, then introduce the performances of models on our private benchmarks as well as their RS scores in Section 4.2 and Section 4.3, respectively.

4.1 Settings

Evaluated Models We categorize the models benchmarked into two groups: open-source LLMs and close AI models.

• **Open-source LLMs**: This category includes open-source LLMs. We evaluate GPT-J-6B (Wang and Komatsuzaki 2021), Phi-1 (Gunasekar et al. 2023), Phi-1.5 (Li et al. 2023), Phi-2 (Javaheripi et al. 2023), Phi-3-Mini-4K-Instruct (Microsoft 2024), Phi-3.5-Mini-Instruct (Microsoft 2024),Phi-4-Reasoning, Phi-4-Reasoning-Plus, Llama-2-7B-Chat-HF (Touvron et al. 2023), Llama-3-8B-Instruct (AI@Meta 2024), Llama-3.1-8B, Llama-3.1-8B-Instruct (AI@Meta 2024), Llama-3.1-70B-Instruct (AI@Meta 2024), Llama-3.1-Nemotron-70B-Instruct-HF (Wang et al. 2024), Llama-3.2-3B, Llama-3.3-70B-Instruct, Qwen2-7B-Instruct (qwe 2024), Qwen2.5-7B (qwe 2024), Qwen2.5-7B-Instruct, Qwen2.5-Math-7B, Qwen2.5-Math-7B-Instruct, Qwen2.5-72B-Instruct (qwe 2024), Qwen3-8B, Qwen3-14B, Qwen3-32B, Yi-1.5-34B-Chat, Kimi-K2, and Deepseek-Chat-V3.

• **Close AIs**: This group consists of LLMs available as commercial services, providing API access for integration into various applications. Our experiments cover GPT-3.5-turbo, GPT-4 (OpenAI 2024a), GPT-4o (OpenAI 2024b), Claude-3.5-Sonnet, Claude-3.7-Sonnet, Claude-4-Sonnet, Gemini-2.0-Flash-001, and Gemini-2.5-Flash.

Implementation Details We assess the performance of the aforementioned LLMs with LM Evaluation Harness (Gao et al. 2024). Following previous studies (Shang et al. 2025; Bai et al. 2025; Liang et al. 2025a; Wang 2024; Xiao et al. 2025; Liang et al. 2025b), we use greedy search in the generation, with the maximized token length of 50. We use the "exact matching" for seeking answers and compute the accuracy among all samples. While the dataset covers eight different domains, the generation process remains consistent across them. All open-source LLMs are executed with $4 \times$ Nvidia H100 GPUs.

4.2 Evaluating The Performances

We conduct experiments on our private benchmarks, ArxivRollBench2024b with Sequencing (S), Cloze (C), and Prediction (P) among 8 domains, where the results on sequencing are shown in Table 2. We identify several key findings:

• **Open-source LLMs show performance comparable to closeAIs.** Open-source LLMs have shown remarkable

Model name	ArxivRollBench-2024b (S)							
	CS	Q-Fin.	Math	Phy.	Stat.	Bio.	Econ.	EESS
GPT-J-6B	10.3 ± 0.6	12.2 ± 1.1	8.0 ± 0.6	11.7 ± 0.7	9.7 ± 0.5	12.0 ± 0.8	9.7 ± 1.0	12.5 ± 0.5
Phi-1	5.6 ± 0.4	6.9 ± 0.9	7.2 ± 0.6	7.5 ± 0.6	7.6 ± 0.4	5.1 ± 0.6	6.8 ± 0.9	6.3 ± 0.4
Phi-1.5	22.7 ± 0.8	20.9 ± 1.4	25.2 ± 0.9	23.9 ± 1.0	22.5 ± 0.7	23.6 ± 1.1	24.5 ± 1.5	21.4 ± 0.7
Phi-2	23.2 ± 0.8	22.8 ± 1.4	24.8 ± 0.9	24.4 ± 1.0	23.6 ± 0.7	24.2 ± 1.1	24.8 ± 1.5	23.1 ± 0.7
Phi-3-Mini-4K-Instruct	6.3 ± 0.4	4.8 ± 0.7	5.3 ± 0.5	3.4 ± 0.4	6.7 ± 0.4	5.3 ± 0.6	5.1 ± 0.7	6.4 ± 0.4
Phi-3.5-Mini-Instruct	19.8 ± 0.7	20.3 ± 1.4	19.2 ± 0.9	17.9 ± 0.9	19.1 ± 0.7	18.6 ± 1.0	19.3 ± 1.3	19.4 ± 0.6
Phi4-Reasoning	2.0 ± 0.3	2.2 ± 0.5	3.4 ± 0.4	1.3 ± 0.3	1.9 ± 0.2	1.9 ± 0.4	2.5 ± 0.5	1.4 ± 0.2
Phi4-Reasoning-Plus	11.1 ± 0.6	13.1 ± 1.2	10.9 ± 0.7	9.0 ± 0.6	9.8 ± 0.5	10.6 ± 0.8	13.0 ± 1.1	9.2 ± 0.5
Qwen2-7B-Instruct	26.6 ± 0.8	27.9 ± 1.5	25.7 ± 1.0	26.4 ± 1.0	28.3 ± 0.8	27.0 ± 1.2	27.9 ± 1.5	27.6 ± 0.7
Qwen2.5-7B	23.7 ± 0.8	24.8 ± 1.5	22.1 ± 0.9	23.9 ± 1.0	23.4 ± 0.7	26.8 ± 1.1	25.3 ± 1.5	24.3 ± 0.7
Qwen2.5-7B-Instruct	27.6 ± 0.8	26.5 ± 1.5	28.6 ± 1.0	28.3 ± 1.0	26.7 ± 0.7	28.2 ± 1.2	28.3 ± 1.5	27.4 ± 0.7
Qwen2.5-Math-7B	16.7 ± 0.7	18.4 ± 1.3	18.8 ± 0.9	17.7 ± 0.9	17.5 ± 0.6	17.0 ± 1.0	17.2 ± 1.3	15.6 ± 0.6
Qwen2.5-Math-7B-Instruct	5.0 ± 0.4	4.7 ± 0.7	3.7 ± 0.4	3.6 ± 0.4	6.7 ± 0.4	4.9 ± 0.6	7.4 ± 0.9	6.0 ± 0.4
Qwen2.5-72B-Instruct	20.5 ± 0.7	21.8 ± 1.4	17.8 ± 0.8	18.6 ± 0.9	18.8 ± 0.7	22.1 ± 1.1	18.3 ± 1.3	21.8 ± 0.7
Qwen3-8B	31.0 ± 0.9	31.3 ± 1.6	29.0 ± 1.0	28.7 ± 1.0	30.3 ± 0.8	28.5 ± 1.2	27.5 ± 1.5	29.2 ± 0.7
Qwen3-14B	4.7 ± 0.4	6.0 ± 0.8	6.3 ± 0.5	5.0 ± 0.5	5.4 ± 0.4	5.1 ± 0.6	5.1 ± 0.7	4.9 ± 0.4
Qwen3-32B	20.2 ± 0.7	22.2 ± 1.4	20.7 ± 0.9	17.8 ± 0.9	20.2 ± 0.7	19.9 ± 1.0	18.3 ± 1.3	20.1 ± 0.7
Llama2-7B-Chat-HF	7.5 ± 0.5	8.5 ± 1.0	10.0 ± 0.7	6.3 ± 0.5	7.8 ± 0.5	7.3 ± 0.7	10.4 ± 1.0	6.8 ± 0.4
Llama3-8B	22.9 ± 0.8	22.8 ± 1.4	21.7 ± 0.9	23.0 ± 0.9	22.3 ± 0.7	23.6 ± 1.1	20.5 ± 1.4	21.4 ± 0.7
Llama3.1-8B	26.0 ± 0.8	24.2 ± 1.5	24.4 ± 0.9	25.3 ± 1.0	24.7 ± 0.7	25.3 ± 1.1	21.3 ± 1.4	23.0 ± 0.7
Llama3.1-8B-Instruct	28.5 ± 0.8	25.2 ± 1.5	28.6 ± 1.0	27.4 ± 1.0	26.8 ± 0.8	26.1 ± 1.1	24.9 ± 1.5	25.5 ± 0.7
Llama3.1-70B-Instruct	31.4 ± 0.9	34.0 ± 1.6	29.3 ± 1.0	30.9 ± 1.0	30.3 ± 0.8	33.7 ± 1.2	31.9 ± 1.6	32.2 ± 0.8
Llama3.1-Nemotron-70B	33.3 ± 0.9	35.8 ± 1.6	30.1 ± 1.0	32.8 ± 1.1	32.1 ± 0.8	34.4 ± 1.2	33.2 ± 1.6	34.4 ± 0.8
Llama3.2-1B	24.0 ± 0.8	23.6 ± 1.5	25.8 ± 1.0	25.3 ± 1.0	23.8 ± 0.7	25.0 ± 1.1	26.2 ± 1.5	24.1 ± 0.7
Llama3.2-3B	23.1 ± 0.8	21.1 ± 1.4	19.2 ± 0.9	22.0 ± 0.9	21.6 ± 0.7	23.0 ± 1.1	24.3 ± 1.4	21.4 ± 0.7
Llama3.3-70B-Instruct	37.3 ± 0.9	39.0 ± 1.7	34.9 ± 1.0	36.4 ± 1.1	36.0 ± 0.8	37.7 ± 1.3	37.1 ± 1.6	37.4 ± 0.8
Yi1.5-34B	28.1 ± 0.8	28.1 ± 1.5	25.9 ± 1.0	26.5 ± 1.0	29.8 ± 0.8	27.1 ± 1.2	25.7 ± 1.5	27.9 ± 0.7
Kimi-K2	35.7 ± 7.5	40.8 ± 7.1	50.0 ± 8.7	40.0 ± 7.4	44.4 ± 7.5	41.9 ± 7.6	41.7 ± 7.2	43.8 ± 7.2
Deepseek-Chat-V3	45.2 ± 7.8	38.8 ± 7.0	50.0 ± 8.7	44.4 ± 7.5	42.2 ± 7.4	44.2 ± 7.7	41.7 ± 7.2	50.0 ± 7.3
GPT-3.5-turbo	38.1 ± 7.6	28.6 ± 6.5	50.0 ± 8.7	20.0 ± 6.0	26.7 ± 6.7	34.9 ± 7.4	14.6 ± 5.1	31.3 ± 6.8
GPT-4	42.9 ± 7.7	42.9 ± 7.1	32.4 ± 8.1	37.8 ± 7.3	40.0 ± 7.4	34.9 ± 7.4	37.5 ± 7.1	41.7 ± 7.2
GPT-4o	42.9 ± 7.7	49.0 ± 7.2	35.3 ± 8.3	31.1 ± 7.0	46.7 ± 7.5	41.9 ± 7.6	39.6 ± 7.1	41.7 ± 7.2
Claude-3.5-Sonnet	38.1 ± 7.6	36.7 ± 7.0	26.5 ± 7.7	37.8 ± 7.3	44.4 ± 7.5	37.2 ± 7.5	35.4 ± 7.0	43.8 ± 7.2
Claude-3.7-Sonnet	33.3 ± 7.4	40.8 ± 7.1	20.6 ± 7.0	37.8 ± 7.3	44.4 ± 7.5	30.2 ± 7.1	25.0 ± 6.3	37.5 ± 7.1
Claude-4-Sonnet	57.1 ± 7.7	51.0 ± 7.2	35.3 ± 8.3	31.1 ± 7.0	57.8 ± 7.4	41.9 ± 7.6	35.4 ± 7.0	37.5 ± 7.1
Gemini-2.0-flash-001	40.5 ± 7.7	44.9 ± 7.2	41.2 ± 8.6	40.0 ± 7.4	37.8 ± 7.3	41.9 ± 7.6	45.8 ± 7.3	41.7 ± 7.2
Gemini-2.5-flash	40.5 ± 7.7	59.2 ± 7.1	35.3 ± 8.3	55.7 ± 7.5	60.0 ± 7.4	46.5 ± 7.7	47.9 ± 7.3	43.8 ± 7.2

Table 2: Evaluation results of current popular models on ArxivRollBench2024b for Sequencing Tasks.

progress in recent years. While the performance of many open-source models remains relatively low, certain models rival proprietary counterparts. For instance, Kimi-K2, the best-performing open-source model, consistently achieves accuracy rates exceeding 40%, closely matching Gemini and Claude and even surpassing it in some tasks.

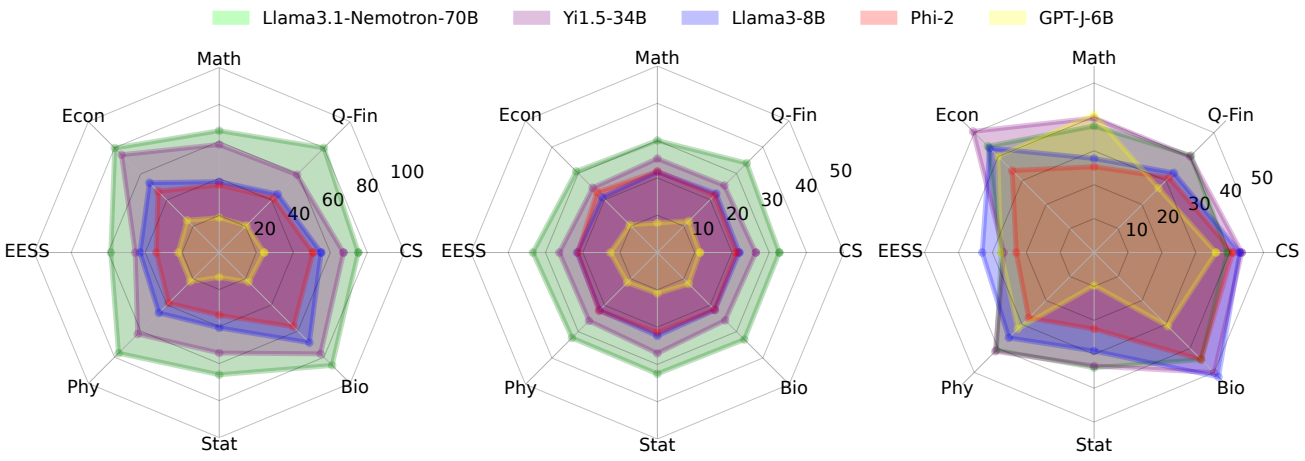
• **Small Language Models (SLMs) are not consistently comparable to medium-sized models.** Table 2 indicates that Phi-3-mini and Phi-3.5-mini perform poorly on Sequencing and Prediction tasks, respectively, with accuracies not exceeding 10%. This suggests that, while SLMs can achieve performance comparable to or even exceeding that of 7-billion-scale models on certain tasks, their actual capabilities may sometimes be over-claimed.

• **While newly emerged LLMs *indeed* achieve better performance, the improvements they claim often reflect *growing* overestimation.** As illustrated in Figure 5, it is evident that within each series, as models evolve, there is an improvement in accuracy. However, the corresponding $\mathbf{RS_I}$ scores also increase on some models (e.g., Phi series). This suggests that while the performance of the models is enhanced through evolution, the degree of overestimation also escalates.

4.3 Evaluating The Overestimation

Analysis on $\mathbf{RS_I}$ and $\mathbf{RS_{II}}$. We provide a detailed comparison of various models in terms of their Absolute $\mathbf{RS_I}$ and Relative $\mathbf{RS_I}$ in Table 3, and $\mathbf{RS_{II}}$ and $\mathbf{N\ RS_{II}}$ in Table 4. Comparing the Absolute $\mathbf{RS_I}$, it is clear that the Qwen and Phi series exhibit the highest degree of overestimation, which are even larger than 100%. Similarly, their corresponding rankings (Relative $\mathbf{RS_I}$) also show significant changes. As for $\mathbf{RS_{II}}$, we observe that models Llama-3.1-Nemotron-70B and Llama3.1-70B score highly on $\mathbf{RS_{II}}$, but their $\mathbf{RS_{II}^N}$ are relatively lower. This discrepancy is due to their high accuracies across various domains in ArxivRollBench and the corresponding high absolute differences. However, after normalization, these differences are not as pronounced.

Measuring Biased Overtraining. We also select five models (GPT-J-6B, Phi2, Llama3-8B, Yi1.5-34B and Llama3.1-Nemotron-70B) as references to analyze their performance across various domain benchmarks and their corresponding Absolute $\mathbf{RS_I}$, as shown in Figure 4. Upon comparison, it is apparent that model performance on public benchmarks is inconsistent, with notably better performance in the domains of Econ, Q-Fin, Bio, and Phy compared to others. However, on ArxivRollBench, the differences in model performance across various domains are minimal, indirectly indicating the



a Performance on public benchmarks b Performance on ArxivRollBench c $\mathbf{RS_I}$ performance

Figure 4: Performance of models across different domain benchmarks and the corresponding Absolute $\mathbf{RS_I}$.

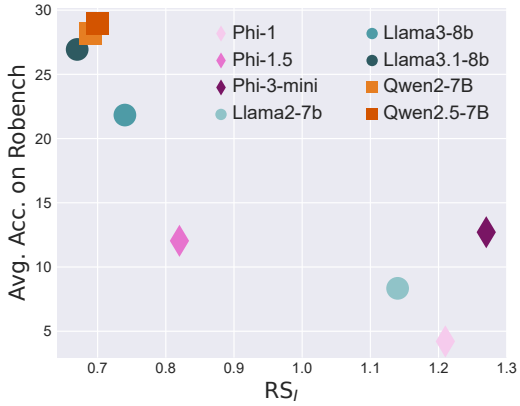


Figure 5: Evolution of series models.

Models	Absolute $\mathbf{RS_I}$	Relative Rank Changes
Phi-1	1.21	↑ 1.31
Phi-1.5	0.82	↓ 0.57
Phi-2	0.62	↓ 0.36
Phi-3-mini	1.27	↑ 0.05
Phi-3.5-mini	1.07	↓ 0.07
Qwen2-7B	0.69	↓ 0.42
Qwen2.5-7B	0.70	↓ 0.37
Qwen2.5-72B	1.41	↓ 0.32
Yi-1.5-34B	0.81	↓ 0.55
Llama-3.1-Nemotron-70B	0.77	↑ 0.14
Llama2-7B	1.14	↑ 0.11
Llama3-8B	0.74	↑ 0.74
Llama3.1-8B	0.67	↑ 0.25
Llama3.1-70B	0.48	↑ 0.02

Table 3: Contamination evaluation with $\mathbf{RS_I}$.

fairness of ArxivRollBench across these domains. Besides, it is observed that Absolute $\mathbf{RS_I}$ are also significantly higher in the Econ, Q-Fin, Bio, and Phy domains, suggesting that advantages of these models on public benchmarks in these areas might be due to overfitting.

5 Conclusion

This paper proposes a novel dynamic evaluation framework called ArxivRoll. It is designed to address the critical issue of overestimation in evaluating LLMs. The framework introduces SCP (Sequencing, Cloze, and Prediction), an automated generator of private test cases, and Rugged Scores (RS), metrics that assess the degree of public benchmark contamination and training bias. Extensive experiments conducted demonstrate the high quality and reliability of the our benchmarks.

Acknowledgements

We would like to express our sincere gratitude to the reviewers for their insightful comments and valuable suggestions.

Model	$\mathbf{RS_{II}}$	N $\mathbf{RS_{II}}$
Phi-1	0.22%	5.21%
Phi-1.5	0.50%	4.02%
Phi-2	0.53%	2.39%
Phi-3-mini	0.76%	5.84%
Phi-3.5-mini	0.57%	3.27%
Qwen2-7B	0.51%	1.76%
Qwen2.5-7B	0.66%	2.21%
Qwen2.5-72B	0.64%	3.84%
Yi-1.5-34B	0.78%	2.91%
Llama-3.1-Nemotron-70B	1.30%	3.88%
Llama2-7B	0.51%	6.20%
Llama3-8B	0.45%	2.00%
Llama3.1-8B	0.96%	3.46%
Llama3.1-70B	1.19%	3.66%

Table 4: Biased overtraining evaluation with $\mathbf{RS_{II}}$.

We are also grateful to Shuxing Fang, Zhenhua Zhou, Yugui Liu, Runshu Wang, Runze Wang, Kunlong Yang, and Junxiang Zhang from The Hong Kong Polytechnic University for their suggestions and assistants, to Yueyue Wang from University of Science and Technology of China for her efforts on this benchmark, and to Professor Hongxia Yang from The Hong Kong Polytechnic University for her constructive feedback. This work was supported by the National Natural Science Foundation of China (Grant No: 92270123 and 62372122), and the Research Grants Council (Grant No: 15210023 and 15207725), and the Innovation and Technology Fund (Grant No: ITS-140-23FP), Hong Kong SAR, China.

References

2024. Qwen2 Technical Report.

Abraham, R. G.; and Chapelle, C. A. 1992. The Meaning of Cloze Test Scores: An Item Difficulty Perspective. *The Modern Language Journal*, 76(4): 468–479.

AI@Meta. 2024. Llama 3 Model Card.

Alderson, J. C. 1979. The Cloze Procedure and Proficiency in English as a Foreign Language. *TESOL Quarterly*, 13(2): 219–227.

Bai, L.; Ye, Q.; Zhang, X.; Zhang, S.; Liang, Z.; Xu, J.; and Hu, H. 2025. Toward Efficient Inference Attacks: Shadow Model Sharing via Mixture-of-Experts. arXiv:2510.13451.

Bormuth, J. R. 1968. The Cloze Readability Procedure. *Elementary English*, 45(4): 429–436.

Britannica. 2024. Gestalt psychology. *Encyclopedia Britannica*.

Chiang, W.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhu, B.; Zhang, H.; Jordan, M. I.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *CoRR*, abs/2110.14168.

Dong, Y.; Jiang, X.; Liu, H.; Jin, Z.; Gu, B.; Yang, M.; and Li, G. 2024. Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 12039–12050. Bangkok, Thailand: Association for Computational Linguistics.

Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Le Noac'h, A.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2024. A framework for few-shot language model evaluation.

Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Giorno, A. D.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; Salim, A.; Shah, S.; Behl, H. S.; Wang, X.; Bubeck, S.; Eldan, R.; Kalai, A. T.; Lee, Y. T.; and Li, Y. 2023. Textbooks Are All You Need. arXiv:2306.11644.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Huang, Z.; Wang, Z.; Xia, S.; Li, X.; Zou, H.; Xu, R.; Fan, R.; Ye, L.; Chern, E.; Ye, Y.; Zhang, Y.; Yang, Y.; Wu, T.; Wang, B.; Sun, S.; Xiao, Y.; Li, Y.; Zhou, F.; Chern, S.; Qin, Y.; Ma, Y.; Su, J.; Liu, Y.; Zheng, Y.; Zhang, S.; Lin, D.; Qiao, Y.; and Liu, P. 2024. OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI. *CoRR*, abs/2406.12753.

Javaheripi, M.; Bubeck, S.; Abdin, M.; Aneja, J.; Bubeck, S.; Mendes, C. C. T.; Chen, W.; Del Giorno, A.; Eldan, R.; Gopi, S.; et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3): 3.

Jiang, M.; Liu, K. Z.; Zhong, M.; Schaeffer, R.; Ouyang, S.; Han, J.; and Koyejo, S. 2024. Investigating Data Contamination for Pre-training Language Models. *CoRR*, abs/2401.06059.

Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. R. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Li, T.; Chiang, W.-L.; Frick, E.; Dunlap, L.; Wu, T.; Zhu, B.; Gonzalez, J. E.; and Stoica, I. 2024a. From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and Bench-Builder Pipeline. *arXiv preprint arXiv:2406.11939*.

Li, T.; Chiang, W.-L.; Frick, E.; Dunlap, L.; Zhu, B.; Gonzalez, J. E.; and Stoica, I. 2024b. From Live Data to High-Quality Benchmarks: The Arena-Hard Pipeline.

Li, Y.; Bubeck, S.; Eldan, R.; Giorno, A. D.; Gunasekar, S.; and Lee, Y. T. 2023. Textbooks Are All You Need II: phi-1.5 technical report. arXiv:2309.05463.

Li, Y.; Guo, Y.; Guerin, F.; and Lin, C. 2024c. An Open-Source Data Contamination Report for Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 528–541. Miami, Florida, USA: Association for Computational Linguistics.

Liang, Z.; Hu, H.; Ye, Q.; Xiao, Y.; and Li, H. 2025a. Why Are My Prompts Leaked? Unraveling Prompt Extraction Threats in Customized Large Language Models. arXiv:2408.02416.

Liang, Z.; Ye, Q.; Wang, Y.; Zhang, S.; Xiao, Y.; Li, R.; Xu, J.; and Hu, H. 2025b. "Yes, My LoRD." Guiding Language Model Extraction with Locality Reinforced Distillation. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1441–1465. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Mather, G. 2006. *Foundations of perception*. Psychology Press.

Microsoft. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219.

Miller, F. 1882. *Telegraphic Code to Insure Privacy and Secrecy in the Transmission of Telegrams*. C.M. Cornwell.

Mirzadeh, S.; Alizadeh, K.; Shahrokhi, H.; Tuzel, O.; Bengio, S.; and Farajtabar, M. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. *CoRR*, abs/2410.05229.

OpenAI. 2024a. GPT-4 Technical Report. arXiv:2303.08774.

OpenAI. 2024b. GPT-4o System Card. arXiv:2410.21276.

Palavalli, M.; Bertsch, A.; and Gormley, M. 2024. A Taxonomy for Data Contamination in Large Language Models. In Sainz, O.; García Ferrero, I.; Agirre, E.; Ander Campos, J.; Jacovi, A.; Elazar, Y.; and Goldberg, Y., eds., *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, 22–40. Bangkok, Thailand: Association for Computational Linguistics.

Shang, H.; Liu, X.; Liang, Z.; Zhang, J.; Hu, H.; and Guo, S. 2025. United Minds or Isolated Agents? Exploring Coordination of LLMs under Cognitive Load Theory. arXiv:2506.06843.

Shannon, C. E. 1949. Communication theory of secrecy systems. *The Bell System Technical Journal*, 28(4): 656–715.

Team, T. T.-B. 2025. Terminal-Bench: A Benchmark for AI Agents in Terminal Environments.

Tech Startups. 2025. Llama 4 Scandal: Meta's release of Llama 4 overshadowed by cheating allegations on AI benchmark.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; and et.al., N. B. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wang, B.; and Komatsuzaki, A. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Wang, Y. 2024. Dehui Du, Haibo Hu, Zi Liang, and Yuanhao Liu. Tsfool: Crafting highlyimperceptible adversarial time series through multi-objective attack. In *European Conference on Artificial Intelligence (ECAI)*.

Wang, Z.; Bukharin, A.; Delalleau, O.; Egert, D.; Shen, G.; Zeng, J.; Kuchaiev, O.; and Dong, Y. 2024. HelpSteer2-Preference: Complementing Ratings with Preferences. arXiv:2410.01257.

White, C.; Dooley, S.; Roberts, M.; Pal, A.; Feuer, B.; Jain, S.; Shwartz-Ziv, R.; Jain, N.; Saifullah, K.; Naidu, S.; Hegde, C.; LeCun, Y.; Goldstein, T.; Neiswanger, W.; and Goldblum, M. 2024. LiveBench: A Challenging, Contamination-Free LLM Benchmark. *CoRR*, abs/2406.19314.

Wu, M.; Zhang, Z.; Dong, Q.; Xi, Z.; Zhao, J.; Jin, S.; Fan, X.; Zhou, Y.; Fu, Y.; Liu, Q.; Zhang, S.; and Zhang, Q. 2025. Reasoning or Memorization? Unreliable Results of Reinforcement Learning Due to Data Contamination. arXiv:2507.10532.

Wu, Z.; Qiu, L.; Ross, A.; Akyürek, E.; Chen, B.; Wang, B.; Kim, N.; Andreas, J.; and Kim, Y. 2024. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. In Duh, K.; Gomez, H.; and Bethard, S., eds., *NAACL*, 1819–1862. Mexico City, Mexico: Association for Computational Linguistics.

Xiao, Y.; Ye, Q.; Hu, L.; Zheng, H.; Hu, H.; Liang, Z.; Li, H.; and Jiao, Y. 2025. Reminiscence Attack on Residuals: Exploiting Approximate Machine Unlearning for Privacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3058–3068.

Xu, C.; Guan, S.; Greene, D.; and Kechadi, M. T. 2024. Benchmark Data Contamination of Large Language Models: A Survey. *CoRR*, abs/2406.04244.

Yang, S.; Chiang, W.; Zheng, L.; Gonzalez, J. E.; and Stoica, I. 2023. Rethinking Benchmark and Contamination for Language Models with Rephrased Samples. *CoRR*, abs/2311.04850.

Zhang, Z.; Chen, J.; and Yang, D. 2024. DARG: Dynamic Evaluation of Large Language Models via Adaptive Reasoning Graph. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhu, K.; Chen, J.; Wang, J.; Gong, N. Z.; Yang, D.; and Xie, X. 2024a. DyVal: Dynamic Evaluation of Large Language Models for Reasoning Tasks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhu, K.; Wang, J.; Zhao, Q.; Xu, R.; and Xie, X. 2024b. Dynamic Evaluation of Large Language Models by Meta Probing Agents. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.