

# How Much Do Large Language Model Cheat on Evaluation? Benchmarking Overestimation under the One-Time-Pad-Based Framework

Zi Liang Liantong Yu Shiyu Zhang Qingqing Ye Haibo Hu

The Hong Kong Polytechnic University

{zi1415926.liang, liantong2001.yu, shiyu187.zhang}@connect.polyu.hk  
{qingqing.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. The source code is available at <https://github.com/liangzid/ArxivRoll/>.

## 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 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<sup>1</sup>, 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, 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<sup>2</sup> for popular LLMs, providing a comprehensive evaluation of their capabilities.

## 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 pre-training of LLMs. Conversely, private benchmarks, namely ArxivRollBench, are created by ArxivRoll and remain confidential until the evaluation period,

<sup>1</sup><https://scale.com/leaderboard>

<sup>2</sup><https://arxivroll.moreoverai.com>

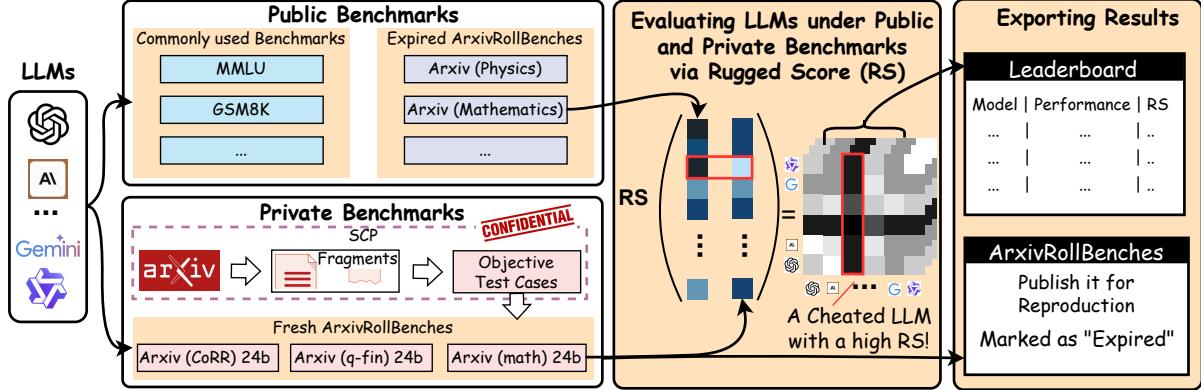


Figure 1: **Framework of ArxivRoll**, which categorizes benchmarks into two distinct groups: public benchmarks and private benchmarks (i.e., ArxivRollBench). These benchmarks are utilized to estimate both the overestimation proportion and performance of large language models (LLMs). Notably, ArxivRoll 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.

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 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<sup>3</sup>, 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

<sup>3</sup><https://arxiv.org/>

A. That is, the output of each sub-layer is  $\text{LayerNorm}(x + \text{Sublayer}(x))$ , where  $\text{Sublayer}(x)$  is the function implemented by the sub-layer itself. B. The encoder is composed of a stack of  $N=6$  identical layers. Each layer has two sub-layers. C. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. We employ a residual connection `\citet{he2016deep}` around each of the two sub-layers, followed by layer normalization `\cite{layernorm2016}`. D. To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of dimension  $\text{dmodel}=512$ .

Selection 1: A C D B Selection 2: B C A D (✓) Selection 3: A B C D Selection 4: D C B A

**Sequencing**

The encoder is composed of a stack of  $N=6$  identical layers. **B** The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. We employ a residual connection `\citet{he2016deep}` around each of the two sub-layers, followed by layer normalization `\cite{layernorm2016}`. **A** To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of dimension  $\text{dmodel}=512$ .

A. That is, the output of each sub-layer is  $\text{LayerNorm}(x + \text{Sublayer}(x))$ , where  $\text{Sublayer}(x)$  is the function implemented by the sub-layer itself.  
B. Each layer has two sub-layers.

Selection 1: A B Selection 2: B A

**Cloze**

The encoder is composed of a stack of  $N=6$  identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. We employ a residual connection `\citet{he2016deep}` around each of the two sub-layers, followed by layer normalization `\cite{layernorm2016}`. **B**

Selection A. In addition to attention sub-layers, each of the layers in our encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically.

Selection B. That is, the output of each sub-layer is  $\text{LayerNorm}(x + \text{Sublayer}(x))$ , where  $\text{Sublayer}(x)$  is the function implemented by the sub-layer itself.

Selection C. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack. Similar to the encoder, we employ residual connections around each of the sub-layers, followed by layer normalization.

**Prediction**

Figure 2: **An illustrative example of symbolic formatting for test samples**, encompassing three formats: sequencing, cloze, and prediction (SCP). We have reformatted SCP into a four-candidate selection task, as detailed in Figure 7.

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 (Appendix C) 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.

### 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  $\mathcal{T} = \{(T_p^i, T_c^i)\}_{i=1,2,\dots,N_p}$  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  $\mathcal{T}_p = \{T_p^j\}_{j=1,2,\dots,N'_p}$  and  $N_c$  unmatchable private benchmarks  $\mathcal{T}_c = \{T_c^k\}_{k=1,2,\dots,N_c}$ , we can define the

rugged score of a model  $m$  as:

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

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 rugged the  $m$ , demonstrating that the evaluated results of  $m$  on public benchmarks  $\{T_p^i\}_{N_p} \cup \{T_p^j\}_{N'_p}$  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 over-training on LLMs, we propose  $\mathbf{RS_{II}}$ , which can be measured by the standard variance on private benchmarks, i.e.,

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

where  $\mathcal{T}_c^p = \{T_c^i, |i = 1, 2, \dots, N_p\}$  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:

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

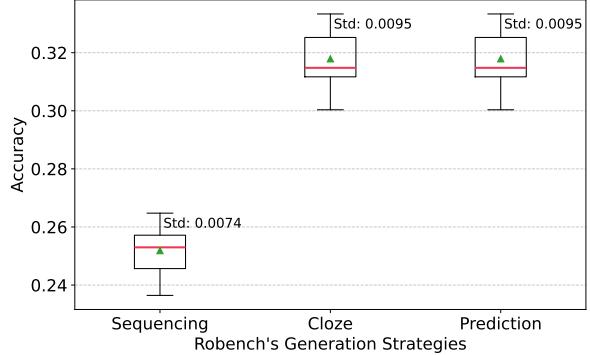


Figure 3: Performance of Llama3 (8B) across 32-time-generated ArxivRollBench benchmarks. The benchmarks were generated 32 times from the same raw article set using SCP. The small standard variance in evaluation results indicates that SCP produces stable test cases.

### 3 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.

#### 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 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 ( $\sim 2.5$  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.

| 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.

### 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

#### 4.1.1 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 (Jawaheripi 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., 2024c), 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.

#### 4.1.2 Implementation Details

We assess the performance of the aforementioned LLMs with LM Evaluation Harness (Gao et al., 2024). Specifically, we use greedy search in the generation, with the maximized token length of

| 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. Results in Cloze and Prediction are in Table 6 and Table 7 in Appendix.

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. Figure 7 in Appendix provides an overview of the instructions we used. All open-source LLMs are executed with 4 × 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, as respectively shown in Table 2, Table 6, and Table 7. We identify several key findings:

- **Open-source LLMs show performance comparable to closeAIs.** Open-source LLMs have shown remarkable progress in recent years. While the performance of many open-source models remains relatively low, certain models rival propri-

etary 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.**

Tables 2 and 7 indicate 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

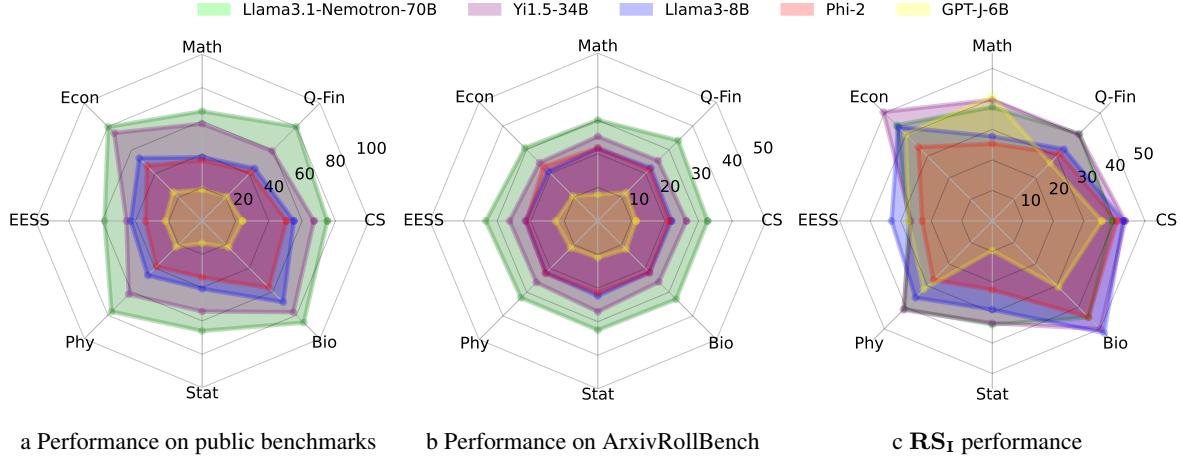


Figure 4: Performance of models across different domain benchmarks and the corresponding Absolute  $\mathbf{RS}_I$ .

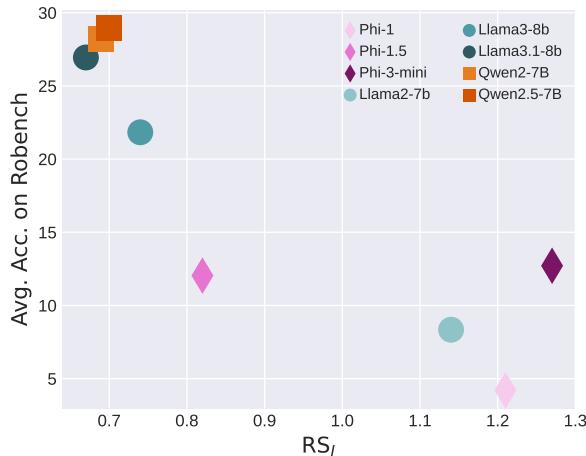


Figure 5: Evolution of series models.

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  $\mathbf{N} \mathbf{RS}_{II}$  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

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

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

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 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.

| Model                  | RS <sub>II</sub> | N RS <sub>II</sub> |
|------------------------|------------------|--------------------|
| 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 RS<sub>II</sub>.

## 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.

## 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.
- 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.
- Elo, A. 1967. The proposed USCF rating system, its development, theory, and applications. *Chess Life* XXII (8): 242–247.
- 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 BenchBuilder 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.
- 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.
- 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.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Biket, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 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, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramanian, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2024a. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; Li, T.; Ku, M.; Wang, K.; Zhuang, A.; Fan, R.; Yue, X.; and Chen, W. 2024b. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. arXiv:2406.01574.
- Wang, Z.; Bukharin, A.; Delalleau, O.; Egert, D.; Shen, G.; Zeng, J.; Kuchaiev, O.; and Dong, Y. 2024c. 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., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1819–1862. Mexico City, Mexico: Association for Computational Linguistics.

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 Contami-  
nation for Language Models with Rephrased Samples.  
*CoRR*, abs/2311.04850.

Ye, T.; Xu, Z.; Li, Y.; and Allen-Zhu, Z. 2024.  
Physics of Language Models: Part 2.1, Grade-  
School Math and the Hidden Reasoning Process.  
arXiv:2407.20311.

Yue, X.; Ni, Y.; Zheng, T.; Zhang, K.; Liu, R.; Zhang,  
G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.;  
Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang,  
Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen,  
W. 2024. MMMU: A Massive Multi-Discipline Mul-  
timodal Understanding and Reasoning Benchmark  
for Expert AGI. In *2024 IEEE/CVF Conference on  
Computer Vision and Pattern Recognition (CVPR)*,  
9556–9567. IEEE.

Zhang, Z.; Chen, J.; and Yang, D. 2024. DARG: Dy-  
namic Evaluation of Large Language Models via  
Adaptive Reasoning Graph. In Globersons, A.;  
Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tom-  
czak, J. M.; and Zhang, C., eds., *Advances in Neural  
Information Processing Systems 38: Annual Confer-  
ence 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, Vi-  
enna, Austria, July 21-27, 2024*. OpenReview.net.

## A A Detailed Description of Our Private Benchmark: ArxivRollBench2024b

**Private Benchmark Details.** As detailed in Section 2.2, we constructed our private benchmarks using preprint papers from Arxiv. Specifically, we downloaded papers uploaded between April 2024 and September 2024 across the following eight domains: Computer Science (CS), Economics (Econ), Electrical Engineering (EESS), Mathematics (Math), Physics (Phy), Biology (Bio), Finance (Fin), and Statistics (Stat). The distribution of collected papers is uneven, with domains such as CS, Math, and Stat containing significantly more articles, while Econ and Fin have fewer. However, even the domains with the smallest number of articles include at least 1,000 papers, ensuring that the constructed benchmark remains robust and suitable for evaluation purposes.

Based on the collected articles, we generated test cases using SCP. Specifically, we began by splitting each article at the “\n” delimiter and then randomly selecting  $N$  consensus phrases to construct a text fragment. To ensure quality, fragments were filtered by checking whether their length exceeded a predefined minimum word count  $N_f$ , eliminating candidates that were too short. Following this, test cases were generated based on SCP as follows:

- Sequencing. The fragment was divided into four parts, permuted, and concatenated.
- Cloze. Four sentences within the fragment were randomly masked.
- Prediction. The last sentence of the fragment was removed, and three similar candidates were retrieved from the article using TF-IDF similarity.

To improve quality, we identified low-quality text samples and developed specific rules to exclude them. Additionally, two annotators manually reviewed the benchmark to further reduce low-quality samples. For the generation process, we set  $N = 1$  and  $N_f = 80$ .

After construction, the final distributions of ArxivRollBench’s private benchmarks are illustrated in Figure 6. Domains such as Econ, Fin, and Bio account for a smaller proportion of samples compared to EESS, CS, Phy, Stat, and Math. A detailed statistical analysis of the private benchmarks is presented in Table 5, revealing that most benchmarks

exhibit high diversity in sequence length. Furthermore, the settings of  $N$  and  $N_f$  result in an average context length of approximately 100 words, making the benchmarks suitable for evaluating most current LLMs.

**Public-Private Benchmark Pairs.** As illustrated in Figure 1, ArxivRollBench leverages both private and public benchmarks to estimate the extent of overestimation. For public benchmarks, we employed two widely used and comprehensive datasets: MMLU and MMLU Pro. Additionally, we supplemented the public benchmarks with domain-specific datasets, such as those commonly used in Math. In the future, expired private benchmarks will also be incorporated into the public benchmarks to further enhance their coverage and utility.

## B External Experiment Results

### C Supplemental Related Work

**Overestimation of LLMs** The overestimation observed during the evaluation of large language models (LLMs) typically stem from two primary sources: *contamination* of test samples and *biased overtraining* on the evaluated tasks.

Contamination in LLMs, as referenced in prior research (Palavalli, Bertsch, and Gormley, 2024; Li et al., 2024c; Dong et al., 2024; Xu et al., 2024; Jiang et al., 2024), occurs when samples in the test dataset have already been included in the pre-training or fine-tuning dataset for a specific LLM. Consequently, the model can achieve superior performance on such a test set not only through task generalization, but also by memorizing the samples. This leads to an unfair and biased comparison among other uncontaminated LLMs, and raises doubts about the actual capabilities of LLMs on specific tasks. Previous studies have demonstrated that test set contamination is a widespread phenomenon in LLMs (Li et al., 2024c; Dong et al., 2024). Furthermore, some methods (Yang et al., 2023) deliberately induce *intentional* contamination to manipulate the benchmark results.

While contamination focuses on cheating at the sample level, another category known as biased overtraining targets the task level to deceive the evaluation. Specifically, when trainers possess prior knowledge about the domains in which their models will be evaluated, they may strategically enhance the performance of their LLMs on these specific domains during pre-training, while neglecting

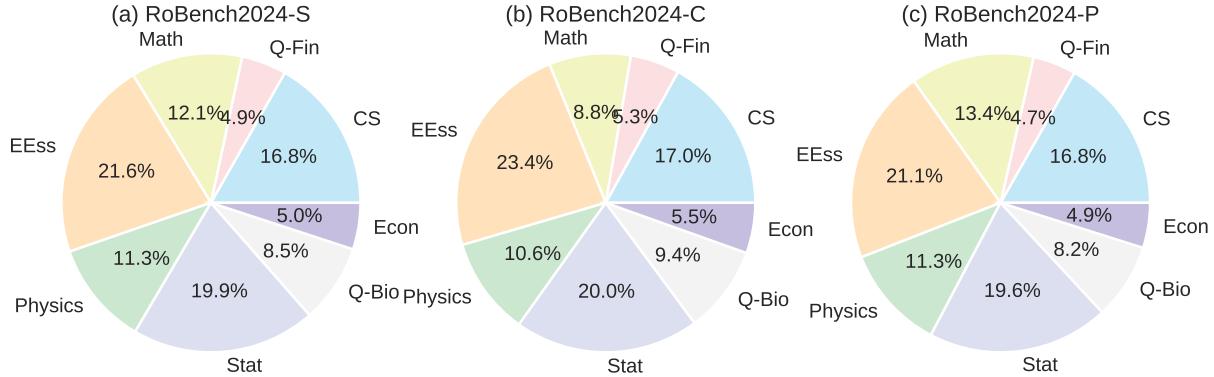


Figure 6: Sample numbers distribution of ArxivRollBench2024b among eight categories across sequencing (a), cloze (b), and prediction (c).

| ArxivRollBench-2024b-S (CS) |           |              |                    |               |               |
|-----------------------------|-----------|--------------|--------------------|---------------|---------------|
| Data Type                   | # Samples | # Avg. Words | # Median Word Num. | Max Word Num. | Min Word Num. |
| Shuffled Text               | 2,931     | 94.89        | 83                 | 612           | 20            |
| ArxivRollBench-2024b-C (CS) |           |              |                    |               |               |
| Data Type                   | # Samples | # Avg. Words | # Median Word Num. | Max Word Num. | Min Word Num. |
| Question Candidates         | 2377      | 117.90       | 102                | 571           | 15            |
| Selections                  | 2377      | 69.15        | 67                 | 342           | 10            |
| ArxivRollBench-2024b-P (CS) |           |              |                    |               |               |
| Data Type                   | # Samples | # Avg. Words | # Median Word Num. | Max Word Num. | Min Word Num. |
| Selections                  | 3166      | 60.76        | 42                 | 542           | 1             |

Table 5: Basic statistical information of ArxivRollBench2024b’s computer science category.

other domains. By doing so, they can manipulate the results of a limited and task-sampled benchmark using their biased and over-trained LLMs. Unlike contamination, biased overtraining has not garnered sufficient attention.

**Robust LLM Evaluation.** To ensure a fair comparison among LLMs and minimize the impact of test sample memorization on a given task, various hand-crafted benchmarks and leaderboards have been proposed. Examples include MMLU-Pro (Wang et al., 2024b), SCI-Bench (Wang et al., 2024a), and MMMU (Yue et al., 2024), which are designed to re-rank the performance of LLMs.

Inspired by these efforts, *private benchmarks* utilizing trusted third-party platforms have emerged as a potential solution to prevent overestimation. However, the assumption of a fully trusted third-party platform is often unrealistic, and the transparency and reproducibility of the evaluation process cannot be guaranteed. Consequently, adversaries may attempt to bribe the platform to improve their ranking or leak test cases for cheating without facing penalties.

Another approach to robust evaluation is the *1-versus-1 arena* (Chiang et al., 2024; Huang et al., 2024; Li et al., 2024b,a), such as Chatbot

Arena (Chiang et al., 2024). In this setup, given the same instruction from a user, the system randomly selects two models, A and B, to answer the question. The user then provides feedback on which model is better. With an infinite number of duels, the ranking of LLMs stabilizes under an elo-based mechanism (Elo, 1967). While this method is effective, it requires a significant amount of evaluation among models and lacks interpretability in terms of why model A is better than model B, both in terms of transparency and reproducibility. Additionally, this leaderboard may be susceptible to malicious annotators who could intentionally provide incorrect feedback.

The third type of robust evaluation focuses on *symbolic formatting* (Mirzadeh et al., 2024; Ye et al., 2024; White et al., 2024; Zhu et al., 2024a; Zhang, Chen, and Yang, 2024; Zhu et al., 2024b). Specifically, for certain tasks, we can design numerous *templates* with placeholders. By flexibly combining these templates and filling in the entities, we can generate an infinite number of test samples. However, this method is only suitable for specific tasks, such as mathematical reasoning, and may be challenging to apply to others, such as commonsense QA and translation.

| Model name               | ArxivRollBench-2024b (C) |             |             |            |            |            |            |            |
|--------------------------|--------------------------|-------------|-------------|------------|------------|------------|------------|------------|
|                          | CS                       | Q-Fin.      | Math        | Phy.       | Stat.      | Bio.       | Econ.      | EESS       |
| GPT-J-6B                 | 2.6 ± 0.3                | 3.3 ± 0.7   | 2.1 ± 0.4   | 4.6 ± 0.5  | 3.5 ± 0.3  | 2.9 ± 0.5  | 2.4 ± 0.5  | 3.5 ± 0.3  |
| Phi-1                    | 1.9 ± 0.3                | 1.9 ± 0.5   | 1.7 ± 0.4   | 1.6 ± 0.3  | 1.9 ± 0.3  | 1.2 ± 0.3  | 1.8 ± 0.5  | 1.6 ± 0.2  |
| Phi-1.5                  | 10.0 ± 0.6               | 11.6 ± 1.2  | 10.4 ± 0.9  | 9.0 ± 0.7  | 13.5 ± 0.6 | 10.7 ± 0.9 | 11.8 ± 1.2 | 10.8 ± 0.5 |
| Phi-2                    | 16.9 ± 0.8               | 19.1 ± 1.4  | 19.0 ± 1.1  | 19.1 ± 1.0 | 18.2 ± 0.7 | 15.7 ± 1.0 | 19.4 ± 1.4 | 18.1 ± 0.7 |
| Phi-3-Mini-4K-Instruct   | 21.9 ± 0.8               | 22.0 ± 1.5  | 17.3 ± 1.1  | 19.0 ± 1.0 | 20.6 ± 0.8 | 19.5 ± 1.1 | 19.5 ± 1.4 | 22.3 ± 0.7 |
| Phi-3.5-Mini-Instruct    | 27.6 ± 0.9               | 27.8 ± 1.6  | 27.1 ± 1.3  | 30.0 ± 1.2 | 27.3 ± 0.8 | 26.3 ± 1.2 | 24.5 ± 1.6 | 28.8 ± 0.8 |
| Phi4-Reasoning           | 10.9 ± 0.6               | 15.1 ± 1.3  | 15.3 ± 1.0  | 11.7 ± 0.8 | 13.5 ± 0.6 | 11.2 ± 0.9 | 15.6 ± 1.3 | 12.1 ± 0.6 |
| Phi4-Reasoning-Plus      | 23.1 ± 0.9               | 24.1 ± 1.6  | 21.2 ± 1.2  | 19.9 ± 1.0 | 23.4 ± 0.8 | 21.9 ± 1.1 | 21.7 ± 1.5 | 24.0 ± 0.7 |
| Qwen2-7B-Instruct        | 24.3 ± 0.9               | 22.8 ± 1.5  | 27.4 ± 1.3  | 25.4 ± 1.1 | 25.2 ± 0.8 | 23.8 ± 1.2 | 23.7 ± 1.5 | 25.7 ± 0.8 |
| Qwen2.5-7B               | 25.0 ± 0.9               | 24.8 ± 1.6  | 28.2 ± 1.3  | 24.8 ± 1.1 | 28.2 ± 0.9 | 25.9 ± 1.2 | 26.2 ± 1.6 | 27.3 ± 0.8 |
| Qwen2.5-7B-Instruct      | 22.5 ± 0.9               | 22.0 ± 1.5  | 24.8 ± 1.2  | 20.4 ± 1.0 | 23.5 ± 0.8 | 21.2 ± 1.1 | 22.3 ± 1.5 | 24.3 ± 0.7 |
| Qwen2.5-Math-7B          | 24.2 ± 0.9               | 22.4 ± 1.5  | 23.7 ± 1.2  | 24.1 ± 1.1 | 24.8 ± 0.8 | 24.6 ± 1.2 | 24.2 ± 1.6 | 24.9 ± 0.8 |
| Qwen2.5-Math-7B-Instruct | 1.5 ± 0.3                | 1.2 ± 0.4   | 2.3 ± 0.4   | 1.8 ± 0.3  | 1.5 ± 0.2  | 1.2 ± 0.3  | 2.4 ± 0.5  | 1.4 ± 0.2  |
| Qwen2.5-72B-Instruct     | 14.7 ± 0.7               | 15.1 ± 1.3  | 16.3 ± 1.1  | 11.9 ± 0.8 | 16.4 ± 0.7 | 11.2 ± 0.9 | 15.6 ± 1.3 | 12.4 ± 0.6 |
| Qwen3-8B                 | 24.0 ± 0.9               | 23.7 ± 1.6  | 25.3 ± 1.2  | 24.4 ± 1.1 | 24.3 ± 0.8 | 23.9 ± 1.2 | 26.4 ± 1.6 | 24.4 ± 0.8 |
| Qwen3-14B                | 4.3 ± 0.4                | 5.9 ± 0.9   | 7.6 ± 0.8   | 5.7 ± 0.6  | 5.1 ± 0.4  | 4.2 ± 0.6  | 5.9 ± 0.9  | 2.8 ± 0.3  |
| Qwen3-32B                | 4.7 ± 0.4                | 3.7 ± 0.7   | 4.8 ± 0.6   | 3.1 ± 0.5  | 5.5 ± 0.4  | 3.0 ± 0.5  | 3.9 ± 0.7  | 3.7 ± 0.3  |
| Llama2-7B-Chat-HF        | 3.2 ± 0.4                | 1.9 ± 0.5   | 5.4 ± 0.6   | 2.0 ± 0.4  | 2.1 ± 0.3  | 1.2 ± 0.3  | 1.3 ± 0.4  | 1.8 ± 0.2  |
| Llama3-8B                | 18.2 ± 0.8               | 18.2 ± 1.4  | 19.2 ± 1.1  | 18.2 ± 1.0 | 20.4 ± 0.8 | 19.0 ± 1.1 | 17.5 ± 1.4 | 18.6 ± 0.7 |
| Llama3.1-8B              | 14.3 ± 0.7               | 14.2 ± 1.3  | 17.3 ± 1.1  | 15.9 ± 1.0 | 15.6 ± 0.7 | 13.8 ± 1.0 | 13.0 ± 1.2 | 14.9 ± 0.6 |
| Llama3.1-8B-Instruct     | 26.2 ± 0.9               | 23.6 ± 1.6  | 25.2 ± 1.2  | 25.5 ± 1.1 | 25.0 ± 0.8 | 24.5 ± 1.2 | 22.9 ± 1.5 | 27.4 ± 0.8 |
| Llama3.1-70B-Instruct    | 27.2 ± 0.9               | 26.5 ± 1.6  | 25.9 ± 1.2  | 27.1 ± 1.2 | 28.1 ± 0.8 | 27.2 ± 1.2 | 25.9 ± 1.6 | 27.3 ± 0.8 |
| Llama3.1-Nemotron-70B    | 26.9 ± 0.9               | 26.8 ± 1.6  | 25.4 ± 1.2  | 27.5 ± 1.2 | 28.0 ± 0.8 | 27.2 ± 1.2 | 25.5 ± 1.6 | 27.5 ± 0.8 |
| Llama3.2-1B              | 15.4 ± 0.7               | 13.3 ± 1.2  | 12.8 ± 0.9  | 11.5 ± 0.8 | 17.8 ± 0.7 | 13.7 ± 0.9 | 15.4 ± 1.3 | 16.0 ± 0.6 |
| Llama3.2-3B              | 23.8 ± 0.9               | 26.5 ± 1.6  | 24.2 ± 1.2  | 25.4 ± 1.1 | 25.8 ± 0.8 | 23.9 ± 1.2 | 25.1 ± 1.6 | 25.9 ± 0.8 |
| Llama3.3-70B-Instruct    | 13.5 ± 0.7               | 13.3 ± 1.2  | 15.5 ± 1.0  | 14.7 ± 0.9 | 15.1 ± 0.7 | 13.2 ± 0.9 | 11.4 ± 1.2 | 14.1 ± 0.6 |
| Yi1.5-34B                | 19.8 ± 0.8               | 21.2 ± 1.5  | 21.0 ± 1.2  | 20.3 ± 1.0 | 19.9 ± 0.8 | 19.8 ± 1.1 | 18.6 ± 1.4 | 20.6 ± 0.7 |
| Kimi-K2                  | 25.8 ± 8.0               | 27.3 ± 6.8  | 26.7 ± 11.8 | 28.6 ± 8.7 | 24.2 ± 7.6 | 32.4 ± 8.1 | 34.9 ± 7.4 | 40.5 ± 7.7 |
| Deepseek-Chat-V3         | 19.4 ± 7.2               | 15.9 ± 5.6  | 33.3 ± 12.6 | 17.9 ± 7.4 | 24.2 ± 7.6 | 14.7 ± 6.2 | 18.6 ± 6.0 | 19.0 ± 6.1 |
| GPT-3.5-turbo            | 29.0 ± 8.3               | 22.72 ± 6.4 | 40.0 ± 13.1 | 21.4 ± 7.9 | 21.2 ± 7.2 | 20.6 ± 7.0 | 34.9 ± 7.4 | 31.0 ± 7.2 |
| GPT-4                    | 16.1 ± 6.7               | 18.2 ± 5.9  | 33.3 ± 12.6 | 14.3 ± 6.7 | 21.2 ± 7.2 | 23.5 ± 7.4 | 27.9 ± 6.9 | 23.8 ± 6.7 |
| GPT-4o                   | 25.8 ± 8.0               | 15.9 ± 5.6  | 26.7 ± 11.8 | 10.7 ± 6.0 | 30.3 ± 8.1 | 23.5 ± 7.4 | 23.3 ± 6.5 | 16.7 ± 5.8 |
| Claude-3.5-Sonnet        | 25.8 ± 8.0               | 22.7 ± 6.4  | 33.3 ± 12.6 | 35.7 ± 9.2 | 33.3 ± 8.3 | 32.4 ± 8.1 | 34.9 ± 7.4 | 28.6 ± 7.1 |
| Claude-3.7-Sonnet        | 19.4 ± 7.2               | 20.5 ± 6.2  | 40.0 ± 13.1 | 17.9 ± 7.4 | 21.2 ± 7.2 | 20.6 ± 7.0 | 34.9 ± 7.4 | 23.8 ± 6.7 |
| Claude-4-Sonnet          | 12.9 ± 6.1               | 9.1 ± 4.4   | 20.0 ± 10.7 | 7.1 ± 5.0  | 27.3 ± 7.9 | 5.9 ± 4.1  | 11.6 ± 4.9 | 14.3 ± 5.5 |
| Claude-4-Opus            | 6.5 ± 4.5                | 2.3 ± 2.3   | 0.0 ± 0.0   | 0.0 ± 0.0  | 12.1 ± 5.8 | 8.8 ± 4.9  | 14.0 ± 5.3 | 4.8 ± 3.3  |
| Gemini-2.0-flash-001     | 19.4 ± 7.2               | 25.0 ± 6.6  | 40.0 ± 13.1 | 28.6 ± 8.7 | 33.3 ± 8.3 | 32.4 ± 8.1 | 23.3 ± 6.5 | 23.8 ± 6.7 |
| Gemini-2.5-flash         | 22.6 ± 7.6               | 13.6 ± 5.2  | 33.3 ± 12.6 | 14.3 ± 6.7 | 33.3 ± 8.3 | 29.4 ± 7.9 | 23.3 ± 6.5 | 31.0 ± 7.2 |

Table 6: Evaluation results of current popular models on ArxivRollBench for Cloze tasks.

In summary, while various robust evaluation methods have been proposed to address the challenges of evaluating LLMs, each has its own limitations. Therefore, it is crucial to continue exploring new and innovative approaches to ensure a fair, transparent, and reproducible evaluation of LLMs.

| Model name               | ArxivRollBench-2024b (P) |            |            |            |            |            |            |            |  |
|--------------------------|--------------------------|------------|------------|------------|------------|------------|------------|------------|--|
|                          | CS                       | Q-Fin.     | Math       | Phy.       | Stat.      | Bio.       | Econ.      | EESS       |  |
| GPT-J-6B                 | 21.5 ± 0.7               | 20.4 ± 1.4 | 13.7 ± 0.7 | 18.1 ± 0.8 | 19.9 ± 0.7 | 21.0 ± 1.0 | 19.6 ± 1.3 | 21.9 ± 0.7 |  |
| Phi-1                    | 4.9 ± 0.4                | 4.4 ± 0.7  | 3.2 ± 0.4  | 3.2 ± 0.4  | 4.1 ± 0.3  | 5.4 ± 0.6  | 4.8 ± 0.7  | 4.5 ± 0.3  |  |
| Phi-1.5                  | 2.0 ± 0.3                | 2.0 ± 0.5  | 2.1 ± 0.3  | 2.6 ± 0.3  | 1.6 ± 0.2  | 2.8 ± 0.4  | 1.4 ± 0.4  | 2.1 ± 0.2  |  |
| Phi-2                    | 23.7 ± 0.8               | 23.3 ± 1.4 | 21.6 ± 0.8 | 22.9 ± 0.9 | 22.4 ± 0.7 | 25.2 ± 1.1 | 24.6 ± 1.4 | 22.8 ± 0.7 |  |
| Phi-3-Mini-4K-Instruct   | 11.9 ± 0.6               | 13.3 ± 1.1 | 12.3 ± 0.7 | 12.8 ± 0.7 | 12.4 ± 0.5 | 13.0 ± 0.9 | 12.4 ± 1.1 | 11.9 ± 0.5 |  |
| Phi-3.5-Mini-Instruct    | 4.0 ± 0.4                | 4.3 ± 0.7  | 5.6 ± 0.5  | 5.3 ± 0.5  | 4.1 ± 0.3  | 4.6 ± 0.5  | 4.0 ± 0.7  | 3.9 ± 0.3  |  |
| Phi4-Reasoning           | 0.4 ± 0.1                | 0.8 ± 0.3  | 0.4 ± 0.1  | 0.8 ± 0.2  | 0.5 ± 0.1  | 0.5 ± 0.2  | 0.5 ± 0.2  | 0.6 ± 0.1  |  |
| Phi4-Reasoning-Plus      | 21.9 ± 0.7               | 22.5 ± 1.4 | 18.6 ± 0.8 | 21.3 ± 0.9 | 22.6 ± 0.7 | 25.5 ± 1.1 | 26.2 ± 1.5 | 24.4 ± 0.7 |  |
| Qwen2-7B-Instruct        | 34.2 ± 0.8               | 32.0 ± 1.6 | 30.2 ± 0.9 | 32.5 ± 1.0 | 32.9 ± 0.8 | 32.2 ± 1.2 | 33.8 ± 1.6 | 33.4 ± 0.8 |  |
| Qwen2.5-7B               | 32.9 ± 0.8               | 30.3 ± 1.5 | 30.5 ± 0.9 | 33.1 ± 1.0 | 32.9 ± 0.8 | 32.3 ± 1.2 | 33.0 ± 1.6 | 33.7 ± 0.7 |  |
| Qwen2.5-7B-Instruct      | 37.8 ± 0.9               | 34.4 ± 1.6 | 34.1 ± 0.9 | 37.7 ± 1.1 | 37.0 ± 0.8 | 39.4 ± 1.2 | 34.7 ± 1.6 | 37.3 ± 0.8 |  |
| Qwen2.5-Math-7B          | 22.0 ± 0.7               | 22.9 ± 1.4 | 20.5 ± 0.8 | 21.5 ± 0.9 | 21.4 ± 0.7 | 22.6 ± 1.1 | 22.3 ± 1.4 | 21.6 ± 0.7 |  |
| Qwen2.5-Math-7B-Instruct | 13.6 ± 0.6               | 18.3 ± 1.3 | 11.1 ± 0.6 | 14.1 ± 0.8 | 14.8 ± 0.6 | 15.9 ± 0.9 | 17.7 ± 1.3 | 13.9 ± 0.5 |  |
| Qwen2.5-72B-Instruct     | 3.4 ± 0.3                | 5.3 ± 0.7  | 6.3 ± 0.5  | 5.6 ± 0.5  | 3.6 ± 0.3  | 4.3 ± 0.5  | 6.4 ± 0.8  | 3.6 ± 0.3  |  |
| Qwen3-8B                 | 31.3 ± 0.8               | 30.6 ± 1.6 | 26.6 ± 0.9 | 30.0 ± 1.0 | 29.8 ± 0.8 | 31.2 ± 1.2 | 28.9 ± 1.5 | 31.1 ± 0.7 |  |
| Qwen3-14B                | 11.8 ± 0.6               | 11.5 ± 1.1 | 14.3 ± 0.7 | 13.0 ± 0.7 | 11.6 ± 0.5 | 8.4 ± 0.7  | 10.7 ± 1.0 | 10.3 ± 0.5 |  |
| Qwen3-32B                | 3.2 ± 0.3                | 2.7 ± 0.5  | 3.7 ± 0.4  | 3.5 ± 0.4  | 3.1 ± 0.3  | 2.5 ± 0.4  | 3.0 ± 0.6  | 2.6 ± 0.3  |  |
| Llama2-7B-Chat-HF        | 13.7 ± 0.6               | 16.0 ± 1.2 | 10.8 ± 0.6 | 14.5 ± 0.8 | 15.2 ± 0.6 | 16.7 ± 1.0 | 15.7 ± 1.2 | 14.4 ± 0.6 |  |
| Llama3-8B                | 24.9 ± 0.8               | 26.1 ± 1.5 | 24.1 ± 0.9 | 24.6 ± 0.9 | 24.3 ± 0.7 | 22.8 ± 1.1 | 24.9 ± 1.4 | 24.7 ± 0.7 |  |
| Llama3.1-8B              | 24.2 ± 0.8               | 25.1 ± 1.5 | 23.5 ± 0.8 | 25.1 ± 0.9 | 24.9 ± 0.7 | 23.5 ± 1.1 | 26.3 ± 1.5 | 23.5 ± 0.7 |  |
| Llama3.1-8B-Instruct     | 29.5 ± 0.8               | 29.3 ± 1.5 | 26.6 ± 0.9 | 31.5 ± 1.0 | 29.9 ± 0.8 | 28.6 ± 1.1 | 28.4 ± 1.5 | 29.7 ± 0.7 |  |
| Llama3.1-70B-Instruct    | 36.3 ± 0.9               | 37.7 ± 1.6 | 33.4 ± 0.9 | 36.0 ± 1.0 | 37.1 ± 0.8 | 36.9 ± 1.2 | 32.0 ± 1.5 | 37.4 ± 0.8 |  |
| Llama3.1-Nemotron-70B    | 37.9 ± 0.9               | 39.2 ± 1.6 | 34.6 ± 1.0 | 36.9 ± 1.0 | 37.6 ± 0.8 | 37.4 ± 1.2 | 33.6 ± 1.6 | 38.3 ± 0.8 |  |
| Llama3.2-1B              | 26.3 ± 0.8               | 26.2 ± 1.5 | 24.9 ± 0.9 | 25.3 ± 0.9 | 25.4 ± 0.7 | 25.6 ± 1.1 | 25.0 ± 1.4 | 24.2 ± 0.7 |  |
| Llama3.2-3B              | 22.5 ± 0.7               | 23.3 ± 1.4 | 23.1 ± 0.8 | 23.2 ± 0.9 | 22.7 ± 0.7 | 20.7 ± 1.0 | 20.8 ± 1.3 | 21.4 ± 0.7 |  |
| Llama3.3-70B-Instruct    | 38.9 ± 0.9               | 38.5 ± 1.6 | 34.3 ± 0.9 | 37.5 ± 1.0 | 38.4 ± 0.8 | 38.1 ± 1.2 | 36.2 ± 1.6 | 38.0 ± 0.8 |  |
| Yi1.5-34B                | 31.3 ± 0.8               | 27.1 ± 1.5 | 28.8 ± 0.9 | 31.0 ± 1.0 | 31.5 ± 0.8 | 30.3 ± 1.2 | 29.4 ± 1.5 | 30.5 ± 0.7 |  |
| Kimi-K2                  | 48.0 ± 7.1               | 52.0 ± 7.1 | 39.2 ± 6.9 | 31.4 ± 6.6 | 44.0 ± 7.1 | 44.9 ± 7.2 | 52.0 ± 7.1 | 47.1 ± 7.1 |  |
| Deepseek-Chat-V3         | 42.0 ± 7.1               | 56.0 ± 7.1 | 39.2 ± 6.9 | 33.3 ± 6.7 | 52.0 ± 7.1 | 49.0 ± 7.2 | 44.0 ± 7.1 | 47.1 ± 7.1 |  |
| GPT-3.5-turbo            | 24.0 ± 6.1               | 40.0 ± 7.0 | 31.4 ± 6.6 | 33.3 ± 6.7 | 38.0 ± 6.9 | 38.8 ± 7.0 | 32.0 ± 6.7 | 25.5 ± 6.2 |  |
| GPT-4                    | 36.0 ± 6.9               | 52.0 ± 7.1 | 43.1 ± 7.0 | 31.4 ± 6.6 | 60.0 ± 7.0 | 51.0 ± 7.2 | 50.0 ± 7.1 | 37.3 ± 6.8 |  |
| GPT-4o                   | 40.0 ± 7.0               | 52.0 ± 7.1 | 39.2 ± 6.9 | 23.5 ± 6.0 | 58.0 ± 7.1 | 51.0 ± 7.2 | 46.0 ± 7.1 | 39.2 ± 6.9 |  |
| Claude-3.5-Sonnet        | 48.0 ± 7.1               | 58.0 ± 7.1 | 49.0 ± 7.1 | 43.1 ± 7.0 | 54.0 ± 7.1 | 57.1 ± 7.1 | 52.0 ± 7.1 | 58.8 ± 7.0 |  |
| Claude-3.7-Sonnet        | 52.0 ± 7.1               | 60.0 ± 7.0 | 45.1 ± 7.0 | 37.3 ± 6.8 | 52.0 ± 7.1 | 57.1 ± 7.1 | 60.0 ± 7.0 | 58.8 ± 7.0 |  |
| Claude-4-Sonnet          | 58.0 ± 7.1               | 66.0 ± 6.8 | 47.1 ± 7.1 | 41.2 ± 7.0 | 62.0 ± 6.9 | 55.1 ± 7.2 | 60.0 ± 7.0 | 58.8 ± 7.0 |  |
| Gemini-2.0-flash-001     | 38.0 ± 6.9               | 56.0 ± 7.1 | 35.3 ± 6.8 | 37.3 ± 6.8 | 50.0 ± 7.1 | 57.1 ± 7.1 | 42.0 ± 7.1 | 49.0 ± 7.1 |  |
| Gemini-2.5-flash         | 54.0 ± 7.1               | 52.0 ± 7.1 | 43.1 ± 7.0 | 37.3 ± 6.8 | 58.0 ± 7.1 | 59.2 ± 7.1 | 60.0 ± 7.0 | 51.0 ± 7.1 |  |

Table 7: Evaluation results of current popular models on ArxivRollBench for Prediction tasks.

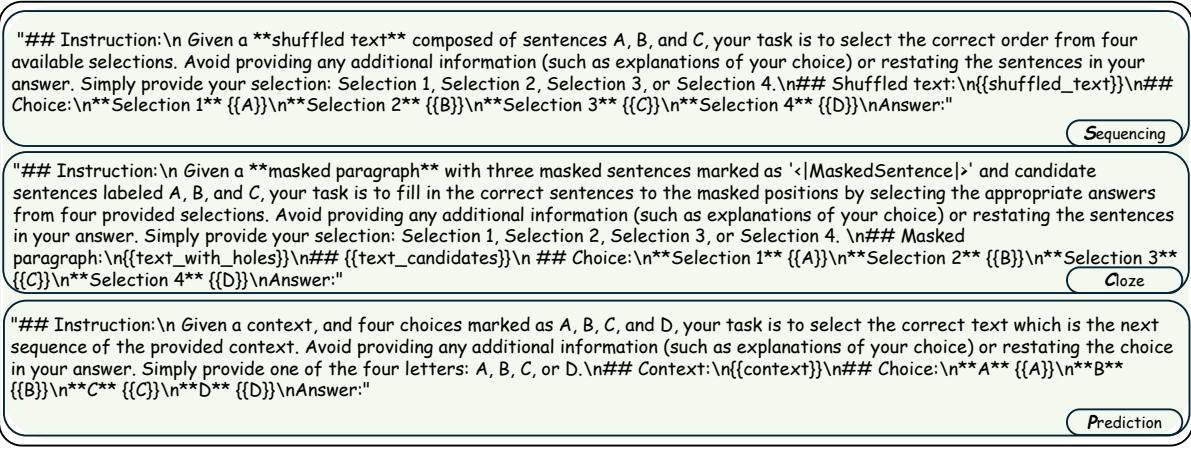


Figure 7: The details of private datasets generation, encompassing three formats: sequencing, cloze, and prediction (SCP).