Data Centric Guard (DC-Guard) - A Framework for Trustworthy LLM Evaluation

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the current era, Large Language Models (LLMs) continue to achieve remarkable results yet their evaluation is increasingly undermined by data-centric challenges such as contamination, memorization and benchmark bias which threaten the reliability of reported performance. To address these issues, we propose DC-Guard (Data Centric Guard), a unified framework for trustworthy evaluation of LLMs. This framework introduces three novel components: the Memorization Consistency Index (MCI) to probe hidden memorization, the Benchmark Ecology Score (BES) to quantify representativeness relative to real-world corpora, and the Contamination-Resilient Metric Adjustment (CRMA) to correct evaluation scores for the contamination risk. Together, these elements provide contamination-aware, bias-adjusted reproducible assessments. Beyond presenting this methodology, we discuss open challenges in maintaining robust evaluations under evolving data sources and shifting usage contexts. DC-Guard offers principled guardrails for fair and transparent benchmarking of the large-scale language models.

1 Introduction

The conventional benchmarks of LLM evaluation suffer from data contamination due to training-test overlap, hidden memorization of surface patterns and coverage bias arising from narrow benchmark distributions. As a result, the existing evaluation practices risk overestimating true model capabilities and undermining reliability. These data centric challenges raise serious concerns about fairness, reproducibility, and interpretability. The community has responded with a variety of methods such as overlap-based contamination detection, quiz-style memorization probes, ad-hoc measures of dataset bias. Yet, these efforts remain fragmented and non-standardized. Existing studies apply different signals, thresholds and definitions, making it difficult to compare results or establish the common ground.

To address this gap, we propose a principled data-centric evaluation framework. The central idea is to treat benchmark evaluation not as a fixed score-reporting exercise, but as an ecological audit of contamination, memorization, and representativeness. This framework unifies scattered techniques into a coherent structure, introducing three new components: the Benchmark Ecology Score (BES) for quantifying coverage bias, the Memorization Consistency Index (MCI) for separating memorization from reasoning, and the Contamination-Resilient Metric Adjustment (CRMA) for reporting fairer accuracy under contamination risk. By reframing evaluation as a structured, contamination-aware ecological process, DC-Guard aims to provide transparent, reproducible, and trustworthy guardrails for both researchers and practitioners.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Literature Survey

The central challenges most of the time revolve around data contamination, memorization, and benchmark representativeness. Several strands of work have emerged addressing these namely,

Data Contamination: One of the earliest and most widely discussed challenge where benchmark items or near-duplicates are already present in the pretraining corpus. This compromises the validity of evaluation since the model may recall the material rather than generalizing it. Approaches for detecting contamination generally fall into two categories:

1. **Surface-level checks:** n-gram overlap or token matching methods, which scan for exact or near-exact text overlap between benchmarks and training corpora. While simple, they fail to capture the semantic rephrasings or paraphrases.
2. **Semantic similarity checks:** Embedding-based methods that measure closeness in the representation space, thereby detecting paraphrased or slightly altered duplicates. These approaches are more robust but still lack a clear calibration for what constitutes meaningful contamination.

Recent studies have introduced more structured tools, such as quiz-style probes that evaluate whether a model can distinguish between original benchmark items and perturbed versions. These methods provide stronger evidence of contamination but remain task-specific and often lack standardization.

Memorization: In this issue, even when evaluation items are not directly present in the training data, LLMs reproduce rare or idiosyncratic information memorized during pretraining. This challenges the notion of "generalization," as models may appear capable when they are in fact recalling. To address this issue, several diagnostic strategies have been proposed such as:

1. **Paraphrase-based Probing:** Checking whether models remain consistent across reworded prompts or equivalent queries. A sharp performance drop under paraphrasing often indicates superficial memorization.
2. **Frequency Analysis:** Investigating whether models disproportionately reproduce sequences that were rare but frequent enough in training to be memorized.
3. **Quiz-style Memorization Detection:** Rephrased or misleading options are introduced to test whether a model is truly reasoning or simply reproducing a memorized pattern.

Despite these advances, memorization detection still struggles with calibration. For instance, if a model answers consistently across paraphrases, is it demonstrating robust reasoning or consistent recall? Current approaches lack a principled metric to separate the two.

Benchmark Representativeness: Another growing concern is the mismatch between benchmarks and real-world usage. Widely used datasets often emphasize narrow domains and stylized problem settings which create a coverage bias where benchmarks may not reflect the diversity of tasks, topics, and linguistic structures encountered in the deployment. Research has sought to quantify representativeness using distributional similarity metrics, comparing benchmarks against large reference corpora. Yet, these studies use different divergence measures and rarely translate their findings into an interpretable score that can be easily adopted with.

In summary, related works have identified the core risks but have addressed them piecemeal. What is missing is a data-centric framework that unifies these strands under a coherent philosophy and provides principled metrics. This motivates our proposal of DC-Guard, which consolidates the fragmented efforts into a single theoretical pipeline.

3 Proposed Framework

We introduce Data Centric Guard (DC-Guard), whose central principle is to reframe benchmark evaluation as an ecological audit. Each benchmark is treated not merely as a static dataset, but as an environment whose validity depends on its freedom from contamination, its resistance to memorization artifacts, and its representativeness of real-world usage.

The DC-Guard is organized into three theoretical pillars, each corresponding to a novel contribution. Benchmark Ecology Score (BES), Memorization Consistency Index (MCI), and Contamination-Resilient Metric Adjustment (CRMA) are linked in a workflow that begins with auditing the dataset, proceeds to auditing the model behavior, and finally yields adjusted evaluation scores that more faithfully represent the true generalization.

1. **Benchmark Ecology Score (BES):** Traditional benchmarks are often narrow in scope and do not reflect the linguistic and topical diversity encountered in the deployment. Hence, BES quantifies the degree of divergence between a benchmark dataset and a real-world reference corpus. The score draws inspiration from ecological diversity indices, where ecosystems are evaluated based on species richness and evenness. Analogously, benchmarks can be viewed as habitats that sample certain linguistic species (topics, styles, reasoning types).

   Computation:

   (a) Represent each benchmark and reference corpus in a shared distributional space, such as topic distributions or sentence embeddings.

   (b) Compute divergence between the two distributions

   (c) Map the divergence to a categorical scale: Low Bias (close alignment), Medium Bias, or High Bias (substantial mismatch).

2. **Memorization Consistency Index (MCI):** It is designed to disentangle memorization from reasoning by measuring, how models respond to paraphrased variants of benchmark prompts while correcting for the background answer frequency.

   Workflow:

   (a) For each benchmark item x, generate paraphrased variants that preserve semantic meaning but alter surface form.

   (b) Compute *Consistency*(x): the fraction of paraphrases where the model produces identical answers.

   (c) Compute *Background Match*(x): the probability that same answer arises in unrelated prompts (capturing rote response patterns).

   (d) Define the index:

   $$\mathrm{MCI}(x) = \mathrm{Consistency}(x) \times \big(1 - \mathrm{BackgroundMatch}(x)\big)$$

   A high MCI typically indicates that responses are driven by memorization rather than reasoning. Medium MCI suggests mixed signals hence, requires closer inspection. Low MCI suggests reliance on reasoning or contextual adaptation rather than rote recall.

3. **Contamination-Resilient Metric Adjustment (CRMA):** This integrates contamination detection and memorization into a single adjusted metric, preventing inflation of performance scores due to training-test overlap. It modifies raw accuracy scores by discounting performance proportional to estimated contamination probability, while integrating the memorization evidence from MCI metric.

   Formulation: Let,

   - Acc = raw accuracy of the model on benchmark items.
   - $\hat{C}$ = calibrated contamination probability, derived from null-model similarity distributions.
   - MCI = Memorization Consistency Index for the same items.

   We define the Adjusted Accuracy as:

   $$Acc_{\text{adj}} = Acc \times \big(1 - \hat{C} \cdot f(MCI)\big) \tag{1}$$

   where $f(MCI)$ scales contamination penalties by memorization evidence. For instance, if contamination is detected but memorization is low, the penalty is reduced, thereby avoiding double penalization. This formulation couples contamination and memorization signals into a single correction factor, ensuring balanced fairness in evaluation.

4 Discussion

While DC-Guard establishes a structured framework for auditing benchmarks and mitigating contamination and memorization, it also opens up several avenues for deeper inquiry. Primarily, the Benchmark Ecology Score (BES) provides a systematic way of quantifying benchmark representativeness by drawing parallels to ecological diversity. It mainly relies on selecting a reference corpus against which diversity is measured. This inevitably introduces bias, which must be carefully addressed. The ecological alignment must be re-assessed periodically as real-world tasks evolve over time. A major challenge is how do we define and update the ground truth ecology in a dynamic language environment.

The Memorization Consistency Index (MCI) uses paraphrased variants and background-match corrections which bridges the gap between anecdotal evidence of memorization and a quantitative diagnostic tool. Paraphrase-based probing has inherent limitation which is generating faithful paraphrases that preserve difficulty, context, and cultural nuance is nontrivial. Overreliance on these automated paraphrasing tools risks in introducing artifacts. Another challenge lies in differentiating productive memorization like recalling factual constants from unproductive memorization like verbatim recall of benchmark items. A nuanced taxonomy of memorization types needs to be developed.

As LLM applications increasingly shift toward interactive, multimodal, and real-time settings, the static text benchmarks alone may prove inadequate. Extending the framework to multi-turn dialogues, multimodal datasets, and task-specific evaluation remains an open frontier. DC-Guard provides three metrics that could serve as standardized reporting tools, improving comparability across studies. But, Standardization itself is difficult: different research groups may operationalize BES or MCI differently depending on corpora, paraphrasing techniques and contamination baselines. Model developers could optimize them without genuinely improving the generalization. Moreover, publicly flagging contaminated benchmarks may inadvertently disincentivize dataset re-use, even when re-use is valid. Balancing transparency with practicality and aligning evaluation standards with policy frameworks, remain as an unresolved societal challenge. Without community-wide guidelines, the results may diverge. There is a need for shared benchmarks, open-source toolkits, and consensus protocols to ensure reproducibility.

5 Conclusion

LLMs have reached unprecedented levels of fluency and task coverage but their evaluation pipelines remain deeply entangled with data centric challenges. Traditional methods often underemphasize these problems, leading to inflated claims of progress and unreliable signals. In this paper, we introduced DC-Guard, a framework that rethinks evaluation from a data-centric perspective.

By combining three dimensions namely Benchmark Ecology Score (BES) to assess benchmark representativeness, Memorization Consistency Index (MCI) to diagnose model recall behavior, and Contamination-Resilient Metric Adjustment (CRMA) to correct inflated scores, the framework provides a holistic approach for auditing the LLM evaluation pipeline. The strength lies not only in its individual components but also in its integration of ecology, memorization, and contamination into a single evaluative lens. This multidimensional view encourages the field to move beyond and achieve more trustworthy measures of generalization.

References

[1] Golchin, S. & Surdeanu, M. (2023) Data Contamination Quiz: A Tool to Detect and Estimate Contamination in Large Language Models. *arXiv preprint arXiv:2311.06233*.

[2] Xu, C., Yan, N., Guan, S., Jin, C., Mei, Y., Guo, Y. & Kechadi, M.-T. (2025) DCR: Quantifying Data Contamination in LLMs Evaluation. *arXiv preprint arXiv:2507.11405*.

[3] Dong, Y., Jiang, X., Liu, H., Jin, Z., Gu, B., Yang, M. & Li, G. (2024) Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models. In *Findings of ACL 2024*, pp. 12039–12050. :contentReference[oaicite:0]index=0

[4] Xu, C. & Yan, N. (2025) TripleFact: Defending Data Contamination in the Evaluation of LLM-driven Fake News Detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 8808–8823. :contentReference[oaicite:1]index=1

[5] Singh, A.K., Kocyigit, M.Y., Poulton, A., Esiobu, D., Lomeli, M., Szilvasy, G. & Hupkes, D. (2024) Evaluation Data Contamination in LLMs: How Do We Measure It and (When) Does It Matter? *arXiv preprint arXiv:2411.03923*.

[6] Xu, C., Guan, S., Greene, D. & Kechadi, M.-T. (2024) Benchmark Data Contamination of Large Language Models: A Survey. *arXiv preprint arXiv:2406.04244*.

[7] Palavalli, M., Bertsch, A. & Gormley, M.R. (2024) A Taxonomy for Data Contamination in Large Language Models. In *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pp. 22–40. :contentReference[oaicite:2]index=2

[8] Nielsen, A.L. & Jordan, M.I. (2022) Statistical Calibration of Semantic Similarity Scores. *Journal of Machine Learning Research*, **23**(140):1–32.

[9] Murphy, N.C., Kulkarni, S. & Haas, P.J. (2023) Domain Representativeness in Language Understanding Benchmarks. In *Findings of ACL*, pp. 987–1000.

Supplementary Material

The lifecycle of an LLM typically unfolds across three stages: pretraining, where the model learns statistical regularities from massive text corpora; fine-tuning and alignment, where the model is adapted to follow instructions or domain-specific data; and evaluation, where performance is reported on benchmarks meant to represent real-world use. Research on evaluation of LLMs has increasingly shifted towards identifying the flaws in the standard benchmarking practices.

Let us view the usage of the theoretical definitions of DC-Guard Metrics namely Benchmark Ecology Score (BES), Memorization Consistency Index (MCI) and Contamination-Resilient Metric Adjustment (CRMA) on widely adopted benchmarks:

Example 1

Let us consider evaluating a model on the widely used SQuAD (Stanford Question Answering Dataset) benchmark, with Wikipedia (2023 snapshot) serving as the reference corpus. Suppose the model has been fine-tuned on QA tasks, we are now interested in quantifying how contamination and memorization affect reported accuracy.

Assumptions:

- Raw accuracy (Acc): 85%
- Benchmark Ecology Score (BES): Medium Bias (factual QA coverage, under-represents reasoning/dialogue diversity)
- Contamination probability ($\hat{C}$): 0.25 (25% overlap chance with pretraining corpus)
- Memorization Consistency Index (MCI): 0.70 (high consistency across paraphrases)
- Function $f(MCI) = MCI$ (linear scaling)

Inferences:

1. **BES:** Medium Bias $\Rightarrow$ Benchmark is focused but not fully representative; lacks reasoning breadth.
2. **MCI:** High value indicates the model tends to repeat answers across paraphrases, signaling memorization.
3. **CRMA:**

$$Acc_{\text{adj}} = 0.85 \times \big(1 - 0.25 \times 0.70\big) = 0.85 \times 0.825 = 0.701$$

The adjusted accuracy falls to 70.1%, showing that contamination and memorization materially inflate reported performance.

Example 2

Let us consider evaluating a model on the DROP (Discrepancy in Reading Comprehension) benchmark, which is specifically designed to test reasoning over passages with arithmetic and logical operations. Suppose the dataset has minimal overlap with pretraining corpora (low contamination), and the model is observed to adapt flexibly across paraphrased question variants.

Assumptions:

- Raw accuracy (Acc): 72%
- Benchmark Ecology Score (BES): Medium Bias (divergence detected but with sufficient topical coverage)
- Contamination probability ($\hat{C}$): 0.05 (low, as benchmark items are adversarially generated)
- Memorization Consistency Index (MCI): 0.25 (low, indicating reliance on reasoning rather than rote recall)
- Function $f(MCI) = MCI$ (linear scaling)

Inferences:

1. **BES:** Medium Bias $\Rightarrow$ Benchmark is somewhat representative but not overly narrow.
2. **MCI:** Low value suggests model responses vary across paraphrases in meaningful ways, pointing toward reasoning reliance rather than memorization.
3. **CRMA:**

$$Acc_{\text{adj}} = 0.72 \times \big(1 - 0.05 \times f(0.25)\big) = 0.72 \times (1 - 0.0125) = 0.72 \times 0.9875 = 0.711$$

Hence, the adjusted accuracy remains close to raw accuracy, reflecting that contamination is not materially inflating the scores.

Example 3

Let us evaluate a model on a subset of MMLU (Massive Multitask Language Understanding), where items range from definitional recall to light reasoning. The benchmark is reasonably aligned with real-world knowledge queries (good ecology), but historical public availability of some items induces moderate contamination. Model behavior suggests a mix of recall and reasoning.

Assumptions:

- Raw accuracy (Acc): 78%
- Benchmark Ecology Score (BES): Low Bias (close alignment to reference corpus)
- Calibrated contamination probability ($\hat{C}$): 0.15 (moderate)
- Memorization Consistency Index (MCI): 0.45 (medium; mixed signals)
- Scaling function: $f(MCI) = MCI$ (linear)

Inferences:

1. **BES:** Low Bias $\Rightarrow$ the benchmark is broadly representative; ecology alone would not explain inflated scores.
2. **MCI:** Medium value indicates partial reliance on recall with some genuine generalization.
3. **CRMA:**

   $$Acc_{\text{adj}} = 0.78 \times \big(1 - 0.15 \times f(0.45)\big) = 0.78 \times \big(1 - 0.0675\big) = 0.78 \times 0.9325 = 0.72735 \approx 72.7\%$$

   The adjustment is noticeable (reflecting moderate contamination and mixed memorization) but smaller than in high-contamination cases.

Therefore, the metrics of DC-Guard: BES contextualizes the benchmark scope, MCI highlights possible memorization artifacts, and CRMA yields an adjusted, contamination-resilient performance score that better reflects generalization.