POSITION PAPER

LLM Benchmark Datasets should be Contamination-Resistant

Ali Al-Lawati · **Jason Lucas** · **Dongwon Lee** · **Suhang Wang**
The Pennsylvania State University · Correspondence: aha112@psu.edu

ICML 2026
43rd International Conference on Machine Learning · Seoul

1 The Problem: Benchmarks Are Contaminated

Modern pretraining corpora ingest nearly all available text — so benchmark test sets inevitably **leak into training data**. Scores then reflect **recall, not generalization**, inflating reported accuracy.

45.4% contamination on C-EVAL by 2023

91.8% contamination across multilingual benchmarks

−13% accuracy drop on a clean GSM8K mirror (Mistral)

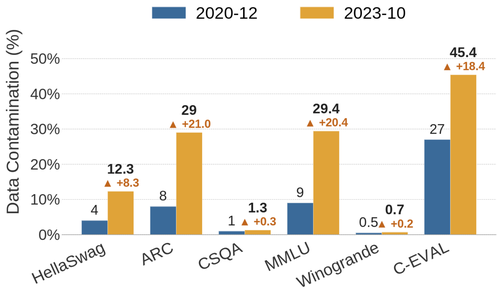

Figure 1. Detected contamination grew sharply from 2020-12 to 2023-10 across popular benchmarks (Li et al., 2024).

Contamination **grows with scale**: larger models memorize verbatim more readily, so even a small contaminated fraction of a trillion-token corpus can compromise evaluation validity.

How leakage happens

- **Direct scraping** — public test sets are copied across repositories, forums, and derivative datasets within months of release.
- **Indirect ingestion** — even gated benchmarks leak via aggregation pipelines, model distillation, and continual pretraining.

Why existing fixes fall short

- **Private benchmarks** — trusted third parties prevent leakage but raise barriers to independent verification.
- **Dynamic benchmarks** — a moving target breaks longitudinal comparison; new data is quickly re-ingested.
- **Rephrasing / decontamination** — degrades difficulty and fails at trillion-token corpus scale.

POSITION STATEMENT

We advocate a fundamental shift toward releasing benchmark datasets in a **contamination-resistant form**: data remains **useful for evaluating models** while being **unlearnable** during public release.

2 Contamination-Resistant Datasets (CRDs)

Definition. A transformed dataset $\varphi(D)$ is contamination-resistant w.r.t. model M if it maintains **inference utility** — $M(\varphi(D))$ yields valid task performance — while being **unlearnable**: gradient steps on $\varphi(D)$ fail to improve generalized performance.

Why not perturb the text? Adversarial-noise tricks fail for discrete text: modern LLMs are robust denoisers, and simple paraphrasing or back-translation strips crafted perturbations. Plaintext must instead be mapped to a **latent form**.

CRD Properties

Irreversibility
Reconstructing the original plaintext from φ(D) is computationally or economically impractical under realistic adversaries.

Equivalence
Model outputs on the CRD approximate outputs on the original data: $M(\varphi(D)) \approx M(D)$.

Interoperability
Given φ(D), a projection $\varphi_1(D)$ can be obtained for any other LLM M_1 that preserves properties 1 & 2.

3 Evaluation Framework

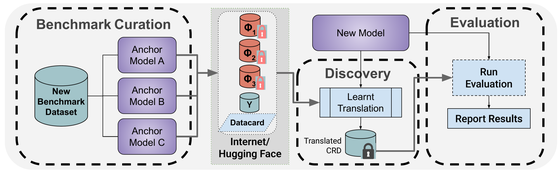

Figure 2. Curation projects the benchmark via anchor models; discovery translates the CRD to the target model; evaluation runs on the translated representation.

A **Curation** — anchor models project questions into latent form; only the inference-required portion is released, plus a datacard and plaintext labels Y.

B **Discovery** — evaluators learn a reusable translation from the anchor's latent space to their target model.

C **Evaluation** — the model generates autoregressively from the translated representation; it never sees plaintext questions. Outputs score against plaintext ground truth (Exact Match, semantic similarity).

Translations are reusable: a learnt target↔anchor mapping serves every future CRD from that anchor — paid once per model, not per benchmark.

4 Key Insight: Training / Inference Asymmetry

Training needs all raw tokens — minimizing next-token loss requires hidden states at every position and layer:

$$L = -\sum_t \log P(x_t \mid x_1, \ldots, x_{t-1})$$

Inference does not: generation continues from only the cached key-value pairs and the penultimate hidden state of the final token — the first new-token query is simply

$$Q_t^{(L)} = h_t^{(L-1)} W_Q^{(L)}$$

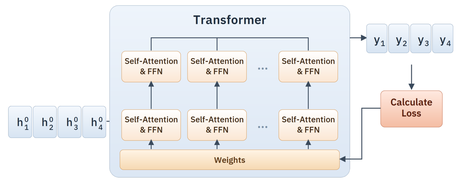

Figure 3a. Training — every raw token must be present to compute per-position loss.

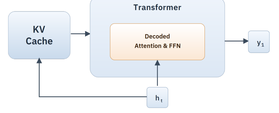

Figure 3b. Inference — KV cache + penultimate hidden state suffice.

Released data-pair format

$$\langle \{K_{1:t}^{(l)}, V_{1:t}^{(l)}\}_{l=1}^{L},\ h_t^{(L-1)},\ Y \rangle$$

Raw tokens are never exposed — the benchmark **cannot be used for pretraining**, yet models still run inference and are scored on plaintext labels Y.

- **Decoding proceeds normally** — each new token is embedded, its KV pair appended to the cache, and attention runs over all cached positions.
- **Fine-tuning is blocked too** — it needs the full raw sequence for label-token hidden states, which the CRD withholds.

Defending Irreversibility

Inversion attacks recover inputs from KV caches mainly under **MHA** — modern **GQA-style models resist**, and lossy representations map many plaintexts to similar embeddings.

- **Layered defenses** — calibrated output noise, entropy-based perturbation, differential privacy, and KV-obfuscation (e.g., KV-Cloak).
- **Stronger settings** — anchor weights can be withheld; encoding offered as a per-model API.

5 Interoperability Across Models

Near-term: anchor models + subspace alignment

Encode benchmarks with widely adopted **anchor models**; translate to targets via closed-form linear maps on dominant singular subspaces (Cross-LoRA-style). Mappings are computed **from weights only** — no plaintext access.

Long-term: relative representations

Project all models into a shared, model-agnostic frame via similarities to **k ≈ 100–500 anchor samples**; angular relations are invariant across latent spaces, enabling zero-shot stitching.

Why this works: the Platonic Representation Hypothesis, CKA, and model stitching show independent networks learn interchangeable representations.

6 Practicality

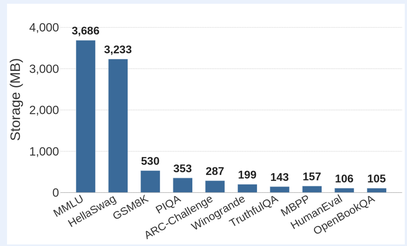

Figure 4. CRD storage per benchmark (Llama2-7B + PyramidKV compression) — MMLU fits in ≈3.7 GB.

- **Storage stays manageable**: retaining just 12% of KV entries preserves performance; ~350 MB per 100K tokens after dropping low-value entries.
- **Compatible** with QA, classification, code, and summarization benchmarks; multi-turn / agentic tasks need adaptation.

Call to Action

- **Advance the CRD ecosystem** — scalable architectures & evaluation methods; test model stitching for low-friction adoption.
- **Standardize anchor models** as universal encoders to cut projection cost for new LLM releases.
- **Integrate CRD validation** into existing pipelines (e.g., Hugging Face) so contamination resistance is seamless.

Supported in part by ARO grant W911NF-21-1-0198 and NSF awards #2114824 and #2438810.