Security and Privacy of Machine Learning, 2025 Critique G11: LLM Memorization – (1) Rethinking LLM Memorization through the Lens of Adversarial Compression; (2) Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data; (3) Memorization Sinks: Isolating Memorization during LLM Training

Shih-Yu Lai
National Taiwan University
Taipei, Taiwan
akinesia112@gmail.com

CRITIQUE:(1) RETHINKING LLM MEMORIZATION THROUGH THE LENS OF ADVERSARIAL COMPRESSION

SUMMARY

The paper proposes **Adversarial Compression Ratio (ACR)** as an operational definition of memorization in LLMs: a string is deemed memorized if a *minimal* prompt, optimized adversarially, is *shorter* than the target string yet elicits it verbatim under greedy decoding Schwarzschild2024ACR. The authors devise MINIPROMPT (built on GCG-style discrete optimization) to approximate the shortest prompt. They validate ACR on four categories (random strings, Associated Press post-cutoff news, Wikipedia, and famous quotes) and show: (i) ICUL (in-context "unlearning") can evade completion-based tests but not ACR; (ii) unlearning on TOFU reduces exact completions but leaves high ACR ("illusion of forgetting"); (iii) LLMs retain Harry-Potter facts post-unlearning; (iv) larger models exhibit higher ACR, consistent with prior memorization scaling trends Carlini2023Quantifying. The work argues ACR is more suitable than completion tests Carlini2023Quantifying, verbatim extraction Nasr2023Scalable, or counterfactual memorization Zhang2023Counterfactual for compliance auditing, and it contrasts with perplexity-as-compression views Deletang2023Language by focusing on *input→output token-count compression* rather than distributional coding.

STRENGTHS

- **Actionable definition.** ACR yields a simple, communicable criterion ("is the minimal prompt shorter than the output?") that is legible to non-ML stakeholders (regulators, counsel), addressing a gap in operational tests Carlini2023Quantifying,Zhang2023Counterfactual.
- **Adversarial robustness to pipeline tweaks.** By optimizing prompts, ACR resists superficial abstention/ICUL layers that defeat naive completion tests; this aligns with realistic, adversarial auditing scenarios.
- **Clear validation design.** The four sanity-check datasets sharply separate memorized vs. unseen/random text; the AP-post-cutoff control is especially persuasive.
- **Revealing unlearning limitations.** The TOFU and Harry Potter case studies demonstrate that reduced completions need not imply reduced *stored* information—a crucial nuance for legal/compliance discussions Nasr2023Scalable.
- **Computational pragmatism.** MINIPROMPT is feasible for model-scale audits (vs. retraining-based counterfactual tests Zhang2023Counterfactual); results are shown across sizes and with a gradient-free baseline to mitigate optimizer idiosyncrasies.

WEAKNESSES

- **Dependence on decoding and tokenization.** ACR is defined under *greedy* decoding; different sampling schemes or minor tokenizer changes may alter minimal prompts and compressibility judgments, complicating standardization.
- **Thresholding and legal calibration.** Using $\tau=1$ (prompt shorter than output) is intuitive but blunt; the SMAZ baseline is informative yet still leaves open domain-, length-, and genre-specific calibration needed for evidentiary use.
- **Exact-match bias.** Results on paraphrases show ACR predominantly captures verbatim memory; this is a feature for copyright, but it under-represents more seman-

tic or template-level memorization that may also raise IP/privacy concerns.
- **Compute and search variance.** While practical, success can hinge on optimizer hyperparameters, search budgets, and random restarts; auditors might reach different ACRs under different budgets, affecting reproducibility and fairness of audits.
- **Scope of models/data.** Most experiments are on transparent open models/datasets; applicability to proprietary SoTA systems (closed weights/data)—the main compliance targets—remains empirically underexplored.

Potential Improvements / Extensions

- **Protocol standardization.** Specify a reference auditing protocol: fixed decoding, tokenizer, search budget, restart count, and stopping rules; report *confidence intervals* for ACR to reflect optimizer variance.
- **Counterfactual baselines without retraining.** Pair ACR with *control models* known not to have seen the target (e.g., pre-cutoff checkpoints) to estimate a *delta-ACR*, strengthening causal claims about training-set exposure.
- **Granular thresholds.** Replace $\tau{=}1$ with *data-dependent* $\tau(y)$ combining SMAZ, target length, and entropy estimates; pre-register domain-specific thresholds (news, code, lyrics) to reduce post-hoc bias.
- **Beyond exact strings.** Add *near-match* ACR variants (e.g., edit-distance or ROUGE-constrained targets) to quantify templatic or partially memorized text, with careful legal framing distinct from verbatim copying.
- **Robustness to guardrails.** Evaluate ACR under layered safety stacks (prompt classifiers, output filters, paraphrase layers) to map where guardrails *actually* reduce compressible memorization vs. only mask completions.
- **Critical domains.** Extend to code, math proofs, and medical snippets where memorization has different stakes than quotes; analyze whether ACR correlates with leakage risk in PII-like contexts.

Questions for Authors

- How stable is ACR across tokenizers and minor model quantization/fine-tuning changes? Could an actor *game* ACR by re-tokenization without materially reducing stored information?
- Would a *stochastic* ACR (measured over temperature/top-p sampling) yield more conservative or more realistic audits than greedy decoding alone?
- Can you formalize conditions under which high ACR *implies* training-set membership with high probability, and bound false positives for non-training text?
- For unlearning evaluation, can *per-sample* ACR trajectories (pre/post each step) predict when gradient-ascent unlearning actually perturbs weight regions storing a string?
- Could delta-ACR across model families/versions serve as a *forensic signal* of data contamination or post-cutoff training?

Critique: (2) Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data

Summary

This paper proposes a scalable methodology to link large language model (LLM) behavior to pretraining data via two constructs: (i) a *task-gram table* mined from supervised input–output pairs by matching semantically similar n-gram pairs; and (ii) a *task-gram language model* that estimates conditional frequencies of those pairs in the pretraining corpus. "Distributional memorization" is operationalized as the Spearman correlation between task-gram LM probabilities and an LLM's token probabilities on test outputs, while "distributional generalization" is defined as the converse. Using Pythia models (13M–12B) trained on The Pile, the authors analyze WMT (translation), TriviaQA (factual QA), MMLU (knowledge vs. reasoning partitions), and GSM8K (math). Findings suggest strong memorization in knowledge-centric TriviaQA, weak/insignificant memorization in WMT and GSM8K, and mixed patterns in MMLU. An influence-function analysis indicates documents containing full n-gram *pairs* have greater training-time impact than documents containing only the output n-gram. Finally, a prompt-rewriting experiment (using an external model and an n-gram-count reward) appears to steer models toward memorization (benefiting TriviaQA) or generalization (benefiting GSM8K). The work situates itself alongside prior studies on verbatim memorization and data attribution, leveraging recent corpus-search infrastructure (e.g., WIMBD, ∞-gram) to operate at larger scales than traditional counterfactual retraining studies [1]–[5].

Strengths

- **Scalable attribution framing.** The task-gram LM neatly bridges task supervision and pretraining statistics without needing costly retraining, advancing beyond classic counterfactual memorization analyses [1].
- **Task-sensitive granularity.** Modeling *paired* n-grams addresses long-range, cross-span dependencies missed by local prefix models such as ∞-gram [4], and resonates with observed template/structure effects in LLM outputs [5].
- **Converging evidence.** Correlational (distributional memorization) and gradient-based influence signals agree directionally: TriviaQA shows the strongest training-data influence; WMT the weakest. This triangulation strengthens the interpretation.
- **Actionable takeaway.** The prompt optimization pilot connects the analysis to practical levers: shaping prompts toward (or away from) pretraining-typical n-gram statistics produces predictable gains on knowledge vs. reasoning tasks.
- **Clear positioning in literature.** The paper relates its construct to verbatim memorization and contamination work [2], and to new pretraining corpus tooling (WIMBD) [3].

Weaknesses

- **Definition asymmetry.** "Generalization" is defined as the inverse of correlation with pretraining frequencies. This collapses multiple qualitatively different behaviors (e.g., systematic compositionality, rule induction, paraphrastic novelty) into a single residual. Alternative divergence measures (e.g., conditional MI, D_{KL}, earth mover's distance on phrase distributions) could disentangle "non-memorization" modes.
- **Sensitivity to mining choices.** Task-gram pairs depend on embedding model, n, and similarity thresholds. Although an ablation is presented, more systematic sensitivity analysis (including alignment noise, polysemy, threshold sweeping with calibration curves) is needed to ensure construct validity.
- **Document-level co-occurrence heuristic.** Counting pair co-occurrence within a document risks spurious associations in heterogeneous sources (e.g., forums, compilations). Windowed co-occurrence or discourse-aware segmentation may reduce false positives.
- **Attribution confounds.** Spearman correlations between two frequency-like objects can be inflated by global popularity, Zipfian effects, or prompt-style artifacts. Controls for unigram/bigram baselines, lexical frequency matching, and answer-length normalization would clarify incremental signal beyond simple frequency.
- **Externalities of prompt tuning.** The optimizer maximizes overlap with pretraining distribution to gain TriviaQA accuracy. This might also *increase* regurgitation/licensing risk in deployment contexts (privacy, copyright), an axis not evaluated in the utility–risk trade-off [2].

Potential Improvements / Extensions

- **Richer distributional metrics.** Replace inverse-correlation as "generalization" with (i) conditional mutual information between task-gram and LLM outputs given input x; (ii) token-level CKA/Centered-Kernel alignment with learned hidden states; or (iii) Wasserstein distance over aligned phrase embeddings.
- **Causal identification beyond influence functions.** Combine gradient-influence with (a) stratified ablations using synthetic proxy corpora; (b) *targeted* pretraining exclusion on narrow slices (feasible at small scale) to calibrate influence estimates; and (c) instrumental variables such as publication-time shocks for certain Wikipedia domains.
- **Windowed, discourse-aware search.** Count pair co-occurrences within sliding windows or sentence/paragraph boundaries; learn window size from data. Compare to TF–IDF-weighted and dependency-arc-aligned pairs to reduce co-location bias.
- **Decomposition of frequency signal.** Orthogonalize pair frequency from (i) unigram frequency, (ii) phrase length, (iii) part-of-speech patterns, and (iv) prompt register. Report partial correlations and Shapley-style attributions to isolate what the *pairing* adds.
- **Cross-corpus generality.** Repeat analyses on newer corpora (e.g., Dolma) and models beyond Pythia; test whether conclusions persist with instruction-tuned checkpoints and contemporary tokenizers.
- **Risk-aware prompt optimization.** Extend the optimizer to a multi-objective RL setting balancing task accuracy, regurgitation rate, and novelty; evaluate across seeds and paraphrase-robust test sets.

Questions for Authors

1) How robust are memorization estimates to alternative n-gram mining schemes (e.g., bilingual lexicons for WMT; span-level aligners for QA) and to different embedding families?
2) Could you report partial correlations controlling for answer length and unigram/TF–IDF frequency to quantify the *incremental* contribution of pair statistics?
3) Influence analysis samples 50 documents per test point. How stable are results under larger R, stratification by domain, and deduplication across near-duplicates?
4) For MMLU's knowledge vs. reasoning split, do misclassified categories (e.g., applied math with domain facts) change the monotonic trends?
5) Prompt optimization improves accuracy; what is its effect on verbatim overlap metrics (e.g., near-duplicate rate) and on privacy-sensitive spans?

Critique: (3) Memorization Sinks: Isolating Memorization during LLM Training

Summary

The paper tackles a central privacy/safety challenge in LLMs: repeated natural-text sequences are memorized and are hard to remove post hoc without collateral damage to model quality [6]. The authors diagnose a key reason—*mechanistic entanglement*—showing theoretically (minimum-norm bias) and empirically (TinyStories, SmolLM 360M/1.7B) that standard training interweaves memorization with generalization, which explains the degradation observed in post-hoc localization methods [7]. They propose **MemSinks**: dedicate a pool of MLP neurons as "sink" units and activate a deterministic subset per sequence via sequence-dependent dropout. This channels sequence-specific signal into sinks while allowing shared neurons to learn generalizable features; dropping sinks at inference time reduces verbatim recall while preserving validation performance. Experiments demonstrate improved forgetting–degradation trade-offs, robustness to modest ID noise, and scalability to billion-token pretraining on SlimPajama with upsampled TinyStories [8]. The analysis connects MemSinks' effectiveness to learning/forgetting cycles documented in prior work [9] and clarifies why generic sparse routing (e.g., MoE) does not automatically yield sequence-level localization [10].

Strengths

- **Paradigm shift.** Recasts unlearning from reactive, post-hoc updates to a proactive *training-time architectural*

prior. This is a crisp and compelling design lens for privacy-by-construction.

- **Clear causal story.** The entanglement diagnosis (theory + controlled data) tightly motivates the method; the link to minimum-norm bias gives a principled explanation for why post-hoc localization degrades capabilities.
- **Thoughtful engineering.** Practical sequence-ID handling (hashing, streamed packing) and tensorized RNG for online masks make the approach implementable at scale.
- **Balanced evaluation.** The paper reports both small-scale ablations (split ratios, activation rate, ID noise) and larger-scale results (360M/1.7B) with the right comparison baselines (standard, dedup, post-hoc localization).
- **Safety framing.** The impact statement acknowledges dual-use concerns and situates MemSinks within responsible data governance.

Weaknesses

- **Metadata dependency.** Reliance on stable sequence IDs limits applicability to messy corpora (mixing, chunking, continual ingestion). The paper shows tolerance to modest noise, but not end-to-end strategies for noisy/unknown IDs.
- **Scope of models/tasks.** Evidence stops at 1.7B pretraining and LM loss. It remains unclear how MemSinks interacts with instruction tuning, RLHF, tool-use, or multi-task mixtures where "repetition" is heterogeneous.
- **Granularity of control.** Sinks are sequence-tied; misuse or mislabeling could inadvertently "hide" useful long-range patterns (e.g., facts recurring across documents) inside sinks and be pruned away.
- **Interpretability of sinks.** The work shows where memorization goes, not *what* is stored. Without mechanistic probes (circuits, features), we cannot assess residual leakage or semantic spillover into shared units.
- **Threat-model coverage.** The evaluation lacks systematic adversarial extraction tests (targeted prompting, paraphrase attacks, sampling-based exfiltration) to quantify real privacy gains beyond loss gaps.

Potential Improvements or Extensions

- **ID-lite routing.** Explore learned or semi-random routing that does not require explicit document IDs: e.g., hash-based token/sketch fingerprints, rolling n-gram hashes, or content-derived anchors robust to packing.
- **MoE hybrids.** Compare MemSinks with MoE-style conditional computation where experts are *regularized* to capture sequence-specific residuals; study whether load-balancing + entropy penalties hinder entanglement [10].
- **Post-training alignment.** Evaluate persistence under SFT/RLHF: do sinks remain isolated after instruction tuning? Introduce a "sink-preservation" regularizer during alignment to prevent re-entanglement.
- **Mechanistic auditing.** Use CKA/CCA similarity, representation probing, and sparse autoencoder techniques to characterize what sinks encode vs. shared layers; release per-layer localization metrics and saliency maps.
- **Privacy stress-tests.** Add membership inference, verbatim extraction under paraphrase/jailbreak prompts, and model editing tests to ground privacy claims in attacker-relevant outcomes.
- **Granular unlearning.** Generalize IDs from sequence-level to (source, topic, entity) tags to support source-level or concept-level unlearning; study compositional masks for overlapping groups.

Questions for the Authors

1) **Stability under downstream tuning:** After SFT/RLHF, do sinks stay isolated or get repurposed? Any signs of re-entanglement without explicit constraints?
2) **Capacity and collisions:** With far fewer sink neurons than sequences, how sensitive is isolation to ID collisions at large scale? Do collisions correlate with residual memorization in shared units?
3) **Bias and fairness:** If repeated data correlate with demographics, could sinks concentrate group-specific patterns that, when dropped, differentially harm performance?
4) **Streaming/continual pretraining:** How do you handle shifting IDs as corpora evolve? Is there a mechanism to "re-key" sinks without catastrophic interference?
5) **Defense-in-depth:** Could MemSinks be combined with deduplication, DP-SGD, or knowledge editing to produce additive privacy gains, or do these interfere with sink isolation?

References

[1] V. Feldman, "Does learning require memorization? a short tale about a long tail," in *STOC*, 2020.

[2] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr, and C. Zhang, "Quantifying memorization across neural language models," in *ICLR*, 2022.

[3] Y. Elazar, A. Bhagia, I. Magnusson, A. Ravichander, D. Schwenk, A. Suhr, P. Walsh, D. Groeneveld, L. Soldaini, J. Dodge *et al.*, "What's in my big data? (wimbd)," 2024.

[4] J. Liu, S. Min, L. Zettlemoyer, Y. Choi, and H. Hajishirzi, "Infini-gram: Scaling unbounded n-gram language models to a trillion tokens," in *First Conference on Language Modeling*, 2024.

[5] W. Merrill, N. A. Smith, and Y. Elazar, "Evaluating n-gram novelty of language models using rusty-dawg," 2024.

[6] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr, and C. Zhang, "Quantifying memorization across neural language models," *arXiv preprint*, 2023.

[7] P. Maini, M. C. Mozer, H. Sedghi, Z. C. Lipton, J. Z. Kolter, and C. Zhang, "Can neural network memorization be localized?" *arXiv preprint arXiv:2307.09542*, 2023.

[8] R. Eldan and Y. Li, "Tinystories: How small can language models be and still speak coherent english?" *arXiv preprint arXiv:2305.07759*, 2023.

[9] M. Toneva, A. Sordoni, R. Tachet des Combes, A. Trischler, Y. Bengio, and G. J. Gordon, "An empirical study of example forgetting during deep neural network learning," *arXiv preprint arXiv:1812.05159*, 2018.

[10] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.