Are Samples Extracted From Large Language Models Memorized?

Chawin Sitawarin [1] **Karan Chadha** [1] **Prasad Buddhavarapu** [1] **John X. Morris** [2 3] **Saeed Mahloujifar** [2]
Chuan Guo [2]

Abstract

Training large language models (LLMs) on diverse datasets, including news, books, and user data, enhances their capabilities but also raises significant privacy and copyright concerns due to their capacity to memorize training data. Current memorization measurements, primarily based on extraction attacks like Discoverable Memorization, focus on an LLM's ability to reproduce training data verbatim when prompted. While various extensions to these methods exist, allowing for different prompt forms and approximate matching, they introduce numerous parameters whose arbitrary selection significantly impacts reported memorization rates. This paper addresses the critical research question of how to compute the *false positive rate* (FPR) of these diverse memorization measurements. We propose a practical definition of FPR and ways to interpret them, offering a more principled approach to select an extraction attack and its parameters. Our findings reveal that while "stronger" extraction attacks often identify more memorized samples, they also tend to have higher FPRs. Notably, some computationally intensive methods exhibit lower extraction rates than simpler baselines when controlling for a fixed FPR.

1. Introduction

The recent quest for new high-quality data sources for training large language models (LLMs) raises an alarming risk of privacy and copyright violations (Tremblay v. OpenAI, Inc., 2023; Kadrey v. Meta Platforms, inc, 2023). These concerns exacerbate as researchers found that LLMs have great capacity at "memorizing" their training data verbatim (Carlini et al., 2021). To estimate such risks, both academia and industry have devised several tools for measuring the memorization phenomenon in LLMs.

Today, most of the memorization measurements are based on *extraction attacks*, and some on *membership inference attacks*. Notably, there is an increase adoption of *Discoverable Memorization* (Carlini et al., 2023) as a memorization measurement scheme (PaLM 2 Team, 2023; Gemini Team, 2024; Gemma Team, 2024a;b; Llama Team, 2024). Under this definition, a 50-token suffix is deemed memorized if the target model generates it verbatim when prompted with the preceding 50-token prefix. Subsequent works extend this by allowing the probing prompt to take multiple forms including a set of small perturbations of the prefix (More et al., 2024) and any prompt shorter than the suffix (Schwarzschild et al., 2024). Some allow a more flexible or approximate matching between the true suffix and the generated texts such as edit distance (Ippolito et al., 2023; Karamolegkou et al., 2023) or probabilistic decoding (Hayes et al., 2024).

These memorization definition come with more parameters to adjust (e.g., prefix and suffix lengths, number of augmentations, number of optimization steps, probability thresholds). When these parameters are relaxed, we capture more memorized samples. However, at the same time, it becomes increasingly difficult to know whether the detected samples are truly memorized or are "false positives," i.e., samples that are flagged as memorized but is not actually memorized. Choices of these parameters are currently arbitrary across industry and academic research, lacking systematic comparisons among them. Needless to say, these choices should not be made lightly as they will significantly affect the final extraction rate which could dictate model releases.

The research question we consider in this work is **how to compute the false positive rate of different memorization measurements**. This problem is ill-defined because, as many past works have also pointed out, we do not have a "ground-truth" or an "oracle" for memorized samples. To demonstrate efficacy of memorization detectors, recent works (Schwarzschild et al., 2024; Hayes et al., 2024) show that the extraction rate on a subset of training set is high whereas the rate on a similar set of non-training samples is low. However, since the evaluated samples are arbitrarily chosen, we cannot cleanly interpret the extraction rates.

[1]Central Applied Science at Meta [2]FAIR at Meta [3]Cornell University. Correspondence to: Chawin Sitawarin <chawin.sitawarin@gmail.come>.

Published at ICML 2025 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Vancouver, Canada. PMLR 267, 2025.

2. False Positives in Memorization Detection

Intuitively, *false positives* should represent training samples of a model that are extracted (i.e., considered memorized by some memorization detection tool) but are not truly memorized by the model. However, as mentioned previously, we also do not have access to an oracle that determines if a given sample is memorized. In this work, we instead turn to a quantity that is well-defined and measurable.

Definition 1 (False Positive Rate of Extraction Attacks)**.** *Given a distribution of samples $\mathcal{D}$, a training set S_n containing n IID samples from D, an a model θ trained on S_n, and S'_m a set of m IID samples from $\mathcal{D}$*

$$TPR_n(\theta, \mathcal{D}) := \frac{1}{n} \sum_{(x_j, y_j) \in S_n} \mathsf{extract}_\theta(x_j, y_j) \quad (1)$$

$$FPR_m(\theta, \mathcal{D}) := \frac{1}{m} \sum_{(x_j, y_j) \in S'_m} \mathsf{extract}_\theta(x_j, y_j) \quad (2)$$

Then, we propose that **one should compare different extraction attacks or memorization detectors by their TPR at a fixed low FPR**, also a popular metric in membership inference literature (Carlini et al., 2022). Intuitively, this metric is reasonable in memorization detection because a good memorization detector should (1) capture more training samples (high TPR) while (2) not falsely flagging non-memorized samples (low FPR). FPR computes the extraction rate on non-training samples which indeed cannot be memorized (assuming no overlap with the training set). In practice, S'_m can be an IID held-out test set. Another interpretation of Definition 1 is to instead view this problem from the membership inference setting where the adversary tries to predict membership of both S_n (as member) and S'_m (as non-member) as accurately as possible.

Synthetic fine-tuning setup. The quantities that we can measure, TPR $= p(\mathsf{extracted} \mid \mathsf{member})$ and FPR $= p(\mathsf{extracted} \mid \neg\mathsf{member})$, are not exactly the same as the quantities we truly desire to measure, $p(\mathsf{extracted} \mid \mathsf{memorized})$ and $p(\mathsf{extracted} \mid \neg\mathsf{memorized})$. This is simply because memorization is generally not equivalent to membership but only a subset. To make an assumption that all training samples are memorized (i.e., member and $\mathsf{extracted}$ are equivalent), **we have to create a synthetic setup where it holds by simply training the model for a large number of epochs.** Here, we can reasonably assume that with enough repetition, a model with sufficient capacity memorizes all its training samples.

3. Design Space of Memorization Detection

We first consider a *document*, e.g., a Wikipedia article or a news article, where the adversary has access to the first 50 tokens (*prefix* x) and wants to extract the following sequence of a certain length (*suffix* y). A "sample" is a concatenation of $x \mid\mid y$. In Discoverable Memorization, a suffix y is deemed memorized if the target model generates y verbatim when prompted with the corresponding prefix x, i.e., $y = \hat{y} := \mathsf{gen}_\theta(x)$ where gen_θ represents a generation function from the target model θ using greedy decoding. Unless stated otherwise, we choose the suffix length $|y| = 50$ as suggested by Carlini et al. (2023) and Nasr et al. (2023).

Generalized memorization definition. Instead of prompting the model only with x, it is natural to consider other sets of prompts $z \in \mathcal{Z}$ (Section 3.1). Instead of only the verbatim match, it makes sense to consider other textual similarity metrics or other kinds of distance function $\mathsf{sim}(f_\theta(z), y)$ where $f_\theta(\cdot)$ is some inference process on θ that is not necessarily a greedy decoding (Section 3.2). Lastly, we may also calibrate this similarity metric by subtracting it with another metric $\mathsf{calib}(x, y)$ (Section 3.3).

Definition 2 (Generalized Memorization Definition)**.** *Given a prefix x and a suffix y, y is memorized by an LLM θ if*

$$\mathsf{score}_\theta(x, y) - \mathsf{calib}(x, y) \geq \tau \quad (3)$$

$$\textit{where} \quad \mathsf{score}_\theta(x, y) := \max_{z \in \mathcal{Z}(x)} \mathsf{sim}\left(f_\theta(z), y\right) \quad (4)$$

for some threshold $\tau \in \mathbb{R}$. We call $\mathsf{score}_\theta(y)$ an extraction score *of y from θ.*

Fig. 1 summarizes this section. All of the design axes are orthogonal; any combination is a valid memorization detector.

3.1. Design Axis 1: Prompt

Prompt augmentation. It is unlikely the case that all prefix tokens are necessary, and some may even provide noisy signals (i.e., overlapping sequence pattern with other samples). The idea behind prompt augmentation is that the model is already *almost* capable of generating the suffix verbatim given the full prefix. So just by "perturbing" the prefix slightly, we should increase the chance of generating the suffix more accurately than just prompting with the full prefix. We experiment with four types of augmentation, the first two proposed by More et al. (2024) and the rest by us.

1. **Truncate.** We prompt the model with different truncated versions of x from the front: $\mathcal{Z}(x) = \{x, x[1:], x[2:], \ldots, x[|x|-1:]\}$.
2. **Mask.** We mask out a random subset of tokens in x. Each token has probability p_m of being masked out. The masked out tokens are either (i) simply dropped (Mask-Drop), (ii) replaced with a pad token (Mask-Pad), or (iii) replaced with a random token (Mask-Rand).
3. **Paraphrase.** We paraphrase the prefix with Dipper (Krishna et al., 2023) and Parrot (Damodaran, 2025).
4. **Few-shot prompting.** To best mimic how the training data are presented to the model, we use few-shot prompting method ("FewShot-s" where s is the number of few-

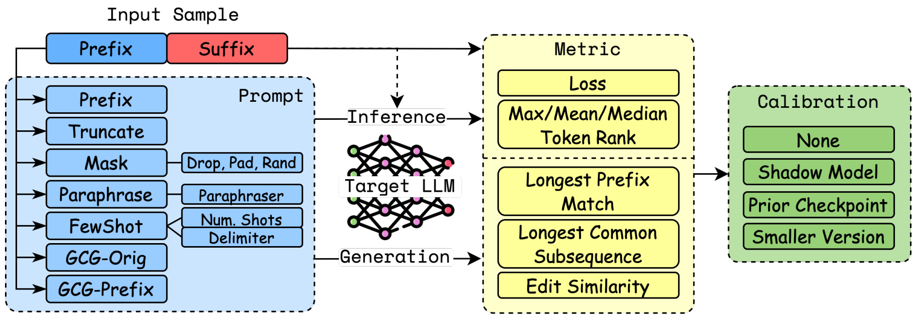

Figure 1: Design space of the extraction-based memorization measurements.

shot samples) that concatenates random training samples from the same dataset with EOS and BOS tokens. We use five few-shot samples by default.

With prompt augmentation, the adversary will prompt the target model multiple times (says N times), and if one of the prompts succeed, we consider the sample memorized.

Prompt optimization. Going beyond prompt augmentation, we naturally expect more sophisticated prompt optimization to further improve the extraction rate. Schwarzschild et al. (2024) propose a memorization definition, termed Adversarial Compression Ratio (ACR), based on a simple form of compression which states that a suffix y is considered memorized if $\exists z$ s.t. $\texttt{gen}(z) = y$ and $|z| \, / \, |x| \leq r$ for some ratio $r \in \mathbb{R}^+$. In other word, we can define $Z = \{z \mid |z| \, / \, |x| \leq r\}$ in Definition 2. ACR uses the GCG attack (Zou et al., 2023), a greedy token-level prompt optimization algorithm originally proposed as jailbreak attacks, to heuristically search the (intractable) prompt space $\mathcal{Z}$.

1. **GCG-Orig**: The prompt is initialized with 50 space-separated "`!`" marks which is done by the original GCG paper (Zou et al., 2023) and Schwarzschild et al. (2024).
2. **GCG-Prefix**: We propose a small variation to GCG-Orig by initializing the prompt with the prefix x.

3.2. Design Axis 2: Metric

Using greedy decoding to generate from the target LLM, we measure how similar the generation is to the true suffix with three popular text similarity metrics: longest prefix match (LPM), longest common subsequence (LCS), and edit similarity (EditSim). Without generating, we can also measure how "close" the model is to generating the true suffix. A natural method is to compute probability of the model generating the true suffix with random decoding $p_\theta(y \mid z)$. We use log of this probability (or negative cross-entropy loss) as the Loss metric.

3.3. Design Axis 3: Calibration

We introduce calibration as a way to combat false positives, borrowing an idea from membership inference literature (Carlini et al., 2021; Watson et al., 2022). The calibration idea is to subtract membership inference score with a measure of "difficulty" of a given sample. This is to help distinguish between memorized members and easy non-members, both of which will incur low loss when not calibrated. Here, we choose a form calibration by a "reference mode" essentially estimating the memorization definition in Feldman & Zhang (2020) with a single reference model, a computationally cheaper version of Zhang et al. (2023). In words, the calibration suggests that if a similarly powerful LLM generates a given suffix without being trained on it, the suffix should *not* be considered memorized even if it is regurgitated by the target model, hence reducing the FPR.

4. Results

4.1. Experiment Setup

Dataset. We use a mixture of three sources published after the training data cutoff date of OLMo-7B. Specifically, we use ArXiv, BBC news, and Wikipedia articles from the RealTimeData[1] dataset (Li et al., 2023) between January 2024 and March 2025. We call this combined dataset $\mathcal{D}_{\mathrm{FT}}$ which we randomly split into $\mathcal{D}_{\mathrm{FT}}^{\mathrm{train}}$, $\mathcal{D}_{\mathrm{FT}}^{\mathrm{test}}$, and $\mathcal{D}_{\mathrm{FT}}^{\mathrm{ref}}$ with a proportion 45%, 45%, and 10%, respectively.

Model. We fine-tune the pre-trained OLMo-7B model (Groeneveld et al., 2024) on $\mathcal{D}_{\mathrm{FT}}^{\mathrm{train}}$ with the next-token prediction objective for 10 epochs to simulate the condition where we believe all training samples are memorized while keeping the memorization detection problem non-trivial. For calibration, we use the pre-trained OLMo-7B as a "prior checkpoint" and also fine-tune OLMo-7B on a smaller $\mathcal{D}_{\mathrm{FT}}^{\mathrm{ref}}$ as "shadow model." For more detailed description and the pre-training setup, please see Appendix C.

4.2. False Positives in Memorization Detection

Both TPR and FPR increase with attack strength. Fig. 2 confirms our hypothesis that stronger attack yields higher extraction rates in both member and non-member sets. Prompting methods that modify prefix more and search over a

[1]`https://huggingface.co/datasets/RealTimeData`

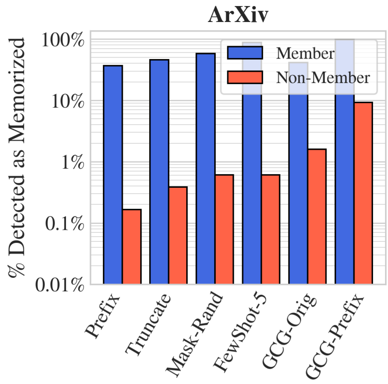

Figure 2: Fraction of samples detected as memorized (LPM $= 1.0$). Both TPR (blue) and FPR (red) generally increase with stronger prompting methods. See Fig. 4 for all data sources.

Table 1: Comparison of different prompting methods on ArXiv under three metrics (LPM, EditSim, and Loss) in the fine-tuning setting. We bold the largest number in each column and underline the second. Prefix prompting is generally a strong choice; Prompt augmentations apart from paraphrase perform better than Prefix, especially FewShot which is on part with (and better at a low FPR) the most expensive GCG-Prefix.

Prompt	AUC (↑)			TPR @ 10% FPR (↑)			TPR @ 1% FPR (↑)		
	LPM	EditSim	Loss	LPM	EditSim	Loss	LPM	EditSim	Loss
Prefix	0.90	0.88	**1.00**	0.78	0.73	**1.00**	0.52	0.42	**0.97**
Truncate	0.93	0.92	**1.00**	0.81	0.81	**1.00**	0.58	0.52	**0.97**
Mask-Rand	0.94	0.94	**1.00**	0.86	0.85	**1.00**	0.64	0.61	**0.97**
Parrot	0.91	0.92	**1.00**	0.78	0.80	**1.00**	0.56	0.52	**0.97**
Dipper	0.91	0.91	**1.00**	0.75	0.78	**1.00**	0.53	0.48	0.94
FewShot-5	**0.99**	**0.99**	0.51	0.98	0.97	0.10	**0.95**	**0.93**	0.01
GCG-Orig	0.75	0.77	0.96	0.56	0.57	0.90	0.39	0.35	0.40
GCG-Prefix	**0.99**	**0.99**	**1.00**	**0.99**	**0.99**	**1.00**	0.00	0.00	0.93

larger space of prompts find more prompts that "extract" a non-member suffix. This raises a question whether all the extracted members are "valid" (i.e., truly memorized by the model) or an artifact of the prompt optimization. When relaxing the metric from $\mathrm{LPM} = 1.0$ to $\mathrm{EditSim} \geq 0.9$ (e.g., from verbatim to approximate match definition used by Ippolito et al. (2023); Gemma Team (2024a;b; 2025)), we observe an expected uptick in both TPRs and FPRs (Fig. 5).

Few-shot prompting has a high TPR but low FPR. FewShot-5 has the second highest TPRs on all three sources, only slightly lower than those of a much more expensive GCG-Prefix, but has much lower FPRs. On the other hand, GCG-Orig is the least promising with relatively high FPRs and low TPR. We will dive into this results in Section 4.3.

Agreement among memorization definitions. we are also interested in how much different memorization definitions agree with one another. Fig. 6 shows a Venn diagram of the training samples with the top-20% scores as determined by each method. It is evident that different prompting methods (Prefix, Mask-Rand, and FewShot-5) find a significant non-overlapping set of memorized samples (more than a third of the samples selected by a given method are not selected by the other two). On the other hand, the three metrics (LPM, EditSim, and Loss) are much more correlated.

4.3. Comparing Memorization Detection

As shown in Section 4.2, different design choices of extract attacks lead to varying TPRs and FPRs. Following the interpretation from Section 2, we now compare different design choices under two popular metrics from membership inference literature: AUC and TPR at a low FPR.

Stronger prompting methods are not necessarily better. Table 1 compares all of the primary prompting methods under three different metrics without calibration. We will highlight results focusing on the LPM metric.

- For generation-based metrics (e.g., LPM and EditSim, but not not Loss), prompt augmentation methods except for paraphrasing improve over the Prefix baseline.
- GCG-Prefix is the best prompting method at high FPR regions (see Fig. 3). However, at low FPRs, it falls short compared to the other non-optimization-based methods. On the other hand, GCG-Orig is overall the worst method across all values of FPRs.
- FewShot-5 performs almost as well as GCG-Prefix at high FPR but much better at low FPR, achieving over 90% TPR at 1% FPR where most methods only reaches 50–70% TPR.

Approximate match metrics. Using the LCS or the EditSim metric does not lead to substantially difference AUC or TPR at a fixed FPR compared to the verbatim match (LPM). This suggests that *on average*, lowering the threshold on LPM is similar to using other more complex approximate match metrics. However, at an *instance level*, predicted positives by LPM and by EditSim do not completely overlap (at a fixed FPR) as mentioned earlier.

Membership inference attacks. Traditional MIA (Prefix + Loss) performs extremely well, beating or on par with all other more sophisticated prompting method. The Loss metric is also better than the generation-based metrics in all cases (except for when used with few-shot prompting).

Calibration. We compare two calibration methods, prior checkpoint and shadow model, in Table 3. In almost all settings, one of the two calibration methods performs better than no calibration. Calibrating the generation-based metric like LPM improves TPRs by a large margin, especially for GCG-Prefix which performs poorly at low FPRs before calibration. For Loss metric, calibration has little effect as TPRs are already close to 100%.

References

Aerni, M., Rando, J., Debenedetti, E., Carlini, N., Ippolito, D., and Tramèr, F. Measuring non-adversarial reproduction of training data in large language models, November 2024. 7

Borkar, J., Jagielski, M., Lee, K., Mireshghallah, N., Smith, D. A., and Choquette-Choo, C. A. Privacy ripple effects from adding or removing personal information in language model training, February 2025. 8

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002. ISSN ISSN 1533-7928. 7

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, Santa Clara, CA, August 2019. USENIX Association. ISBN 978-1-939133-06-9. 7

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. 1, 3, 7, 8

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649. 2

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 7

Chang, H., Shamsabadi, A. S., Katevas, K., Haddadi, H., and Shokri, R. Context-aware membership inference attacks against pre-trained large language models, September 2024. 8

Chang, K., Cramer, M., Soni, S., and Bamman, D. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7312–7327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.453. 8

Chen, J., Wang, J., Peng, T., Sun, Y., Cheng, P., Ji, S., Ma, X., Li, B., and Song, D. Copy, right? A testing framework for copyright protection of deep learning models, December 2021. 8

Damodaran, P. Parrot: Paraphrase generation for NLU, May 2025. 2

Das, D., Zhang, J., and Tramèr, F. Blind baselines beat membership inference attacks for foundation models, June 2024. 8

Duan, J., Kong, F., Wang, S., Shi, X., and Xu, K. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, pp. 8717–8730. PMLR, July 2023. 8

Duarte, A. V., Zhao, X., Oliveira, A. L., and Li, L. DE-COP: Detecting copyrighted content in language models training data, February 2024. 8

Dwork, C. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pp. 1–12, Venice, Italy and Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-35907-9. doi: 10.1007/11787006_1. 7

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2881–2891. Curran Associates, Inc., 2020. 3, 7, 8

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, March 2024. 1

Gemma Team. Gemma: Open models based on gemini research and technology, March 2024a. 1, 4

Gemma Team. Gemma 2: Improving open language models at a practical size, August 2024b. 1, 4

Gemma Team. Gemma 3 technical report. Technical report, 2025. 4

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N. A., and Hajishirzi, H. OLMo: Accelerating the science of language models, June 2024. 3, 8

Hayes, J., Swanberg, M., Chaudhari, H., Yona, I., and Shumailov, I. Measuring memorization through probabilistic discoverable extraction, October 2024. 1, 7

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use, March 2023. 8

Huang, J., Shao, H., and Chang, K. C.-C. Are large pre-trained language models leaking your personal information? In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2038–2047, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.148. 8

Huang, J., Yang, D., and Potts, C. Demystifying verbatim memorization in large language models, July 2024. 7

Ippolito, D., Tramer, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., Choquette Choo, C., and Carlini, N. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In Keet, C. M., Lee, H.-Y., and Zarrieß, S. (eds.), *Proceedings of the 16th International Natural Language Generation Conference*, pp. 28–53, Prague, Czechia, September 2023. Association for Computational Linguistics. 1, 4, 7, 9

Kadrey v. Meta Platforms, inc. Kadrey, et al. v. meta platforms, inc. United States District Court for the Northern District of California, 2023. 1

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference* *on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10697–10707. PMLR, July 2022. 7

Kaneko, M., Ma, Y., Wata, Y., and Okazaki, N. Sampling-based pseudo-likelihood for membership inference attacks, April 2024. 8

Karamolegkou, A., Li, J., Zhou, L., and Søgaard, A. Copyright violations and large language models, October 2023. 1, 7, 8

Kassem, A. M., Mahmoud, O., Mireshghallah, N., Kim, H., Tsvetkov, Y., Choi, Y., Saad, S., and Rana, S. Alpaca against vicuna: Using LLMs to uncover memorization of LLMs, March 2024. 7

Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., and Oh, S. J. ProPILE: Probing privacy leakage in large language models. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023. 7

Krishna, K., Song, Y., Karpinska, M., Wieting, J. F., and Iyyer, M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023. 2

Li, Y., Geurin, F., and Lin, C. LatestEval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *AAAI Conference on Artificial Intelligence*, 2023. doi: 10.48550/arXiv.2312.12343. 3

Liu, K. Z., Choquette-Choo, C. A., Jagielski, M., Kairouz, P., Koyejo, S., Liang, P., and Papernot, N. Language models may verbatim complete text they were not explicitly trained on, March 2025. 8

Llama Team. The llama 3 herd of models, July 2024. 1

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., and Zanella-Beguelin, S. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346–363, Los Alamitos, CA, USA, May 2023. IEEE Computer Society. doi: 10.1109/SP46215.2023.10179300. 8

Ma, I., Domingo, I., Krone-Martins, A., Baldi, P., and Lopes, C. V. Memorization: A close look at books, April 2025. 8

Mattern, J., Mireshghallah, F., Jin, Z., Schoelkopf, B., Sachan, M., and Berg-Kirkpatrick, T. Membership inference attacks against language models via neighbourhood comparison. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.719. 8

Meeus, M., Jain, S., Rei, M., and de Montjoye, Y.-A. Did the neurons read your book? Document-level membership inference for large language models, October 2023. 8

Meeus, M., Shilov, I., Faysse, M., and de Montjoye, Y.-A. Copyright traps for large language models, February 2024. 8

Mireshghallah, F., Uniyal, A., Wang, T., Evans, D., and Berg-Kirkpatrick, T. Memorization in NLP fine-tuning methods, November 2022. 7

More, Y., Ganesh, P., and Farnadi, G. Towards more realistic extraction attacks: An adversarial perspective, July 2024. 1, 2, 7

Nakka, K. K., Frikha, A., Mendes, R., Jiang, X., and Zhou, X. PII-compass: Guiding LLM training data extraction prompts towards the target PII via grounding, July 2024. 8

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models, 2023. 2, 7

Ozdayi, M., Peris, C., FitzGerald, J., Dupuy, C., Majmudar, J., Khan, H., Parikh, R., and Gupta, R. Controlling the extraction of memorized data from large language models via prompt-tuning. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1512–1521, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.129. 7

PaLM 2 Team. PaLM 2 technical report, September 2023. 1

Puerto, H., Gubri, M., Yun, S., and Oh, S. J. Scaling up membership inference: When and how attacks succeed on large language models, October 2024. 8

Schwarzschild, A., Feng, Z., Maini, P., Lipton, Z. C., and Kolter, J. Z. Rethinking LLM memorization through the lens of adversarial compression, April 2024. 1, 3, 7, 8

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(90):2635–2670, 2010. 7

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 8

Su, E., Vellore, A., Chang, A., Mura, R., Nelson, B., Kassianik, P., and Karbasi, A. Extracting memorized training data via decomposition, September 2024. 7

Thudi, A., Shumailov, I., Boenisch, F., and Papernot, N. Bounding membership inference, December 2022. 8

Tiwari, T. and Suh, G. E. Sequence-level leakage risk of training data in large language models, February 2025. 7

Tremblay v. OpenAI, Inc. Tremblay v. OpenAI, Inc. United States District Court, Northern District of California, 2023. 1

Wang, J. G., Wang, J., Li, M., and Neel, S. Pandora's white-box: Precise training data detection and extraction in large language models, May 2024a. 8

Wang, Z., Bao, R., Wu, Y., Taylor, J., Xiao, C., Zheng, F., Jiang, W., Gao, S., and Zhang, Y. Unlocking memorization in large language models with dynamic soft prompting, September 2024b. 7

Watson, L., Guo, C., Cormode, G., and Sablayrolles, A. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*, 2022. 3

Table 2: Summary of the different prompt and scoring methods used in the literature.

Name and Literature	Prompt	Metric	Calibration
Extraction Attack			
Discoverable Memorization (Carlini et al., 2023; Nasr et al., 2023; Huang et al., 2024)	Prefix	LPM	None
Approximate matching (Ippolito et al., 2023; Karamolegkou et al., 2023; More et al., 2024)	Prefix	LCSSeq, EditSim	None
More et al. (2024); Tiwari & Suh (2025)	Truncate	LPM	None
Random token masking (More et al., 2024)	Mask	LPM	None
ACR (Schwarzschild et al., 2024)	GCG-Orig	LPM	None
Membership Inference Attack (MIA)			
Loss MIA (Yeom et al., 2018)	Prefix + Suffix	Loss	None
Carlini et al. (2021)	Prefix + Suffix	Loss	Smaller Version
Mireshghallah et al. (2022)	Prefix + Suffix	Loss	Prior Checkpoint
Counterfactual Memorization (Zhang et al., 2023)	Prefix + Suffix	Loss	Shadow Model

Weller, O., Marone, M., Weir, N., Lawrie, D., Khashabi, D., and Van Durme, B. "According to . . . ": Prompting language models improves quoting from pre-training data. In Graham, Y. and Purver, M. (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2288–2301, St. Julian's, Malta, March 2024. Association for Computational Linguistics. 7

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE, 2018. 7

Yu, W., Pang, T., Liu, Q., Du, C., Kang, B., Huang, Y., Lin, M., and Yan, S. Bag of tricks for training data extraction from language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 40306–40320. PMLR, July 2023. 7

Zhang, C., Ippolito, D., Lee, K., Jagielski, M., Tramèr, F., and Carlini, N. Counterfactual memorization in neural language models. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023. 3, 7

Zhang, J., Das, D., Kamath, G., and Tramèr, F. Membership inference attacks cannot prove that a model was trained on your data, September 2024. 8

Zhao, W., Shao, H., Xu, Z., Duan, S., and Zhang, D. Measuring copyright risks of large language model via partial information probing, September 2024. 7

Zhou, Z., Xiang, J., Chen, C., and Su, S. Quantifying and analyzing entity-level memorization in large language models, November 2023. 8

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models, December 2023. 3, 8

A. Related Work

Memorization definition. Feldman & Zhang (2020) popularized a theoretically appealing definition of memorization in learning algorithm with a strong connections to differential privacy (Dwork, 2006) and algorithmic stability (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010). It is also empirically measured in masked LMs by Zhang et al. (2023). However, this memorization definition is a property of a learning algorithm and not a specific instance of a trained model that we are interested in in practice. Subsequently, a definition of memorization in LMs shifts toward two privacy attacks: training data extraction and membership inference.

Training data extraction. Extraction or reconstruction attacks against LMs were first explored in Carlini et al. (2019) and against autoregressive LLMs in Carlini et al. (2021). Recently, Nasr et al. (2023) measures extraction rates at production scale where LLMs regurgitate their training data verbatim. Subsequent works study multiple factors that affect extraction rates including prompting, sampling methods, and model capacity (Yu et al., 2023; More et al., 2024; Huang et al., 2024; Hayes et al., 2024; Tiwari & Suh, 2025). The most important factor is repetitions of the samples in the training set, and good deduplication has been effective in reducing memorization and also in improving the model performance (Kandpal et al., 2022; Carlini et al., 2023). Prompting method is also another important factor with multiple works exploring extraction via benign conversations (Aerni et al., 2024), prompting with internet texts (Carlini et al., 2021; Nasr et al., 2023), partial information (Weller et al., 2024; Su et al., 2024; Zhao et al., 2024), and adversarial prompting (Kim et al., 2023; Ozdayi et al., 2023; Kassem et al., 2024; Wang et al., 2024b; Schwarzschild et al., 2024).

Different extraction attack determines various different sets of samples as memorized with little consistency. We hypothesize that a fraction of extracted samples are not, in fact, memorized by the target model. We call such samples "false positives," define a way to measure them, and use them to compare a subset of representative extraction attacks from the list above. Our work is most related to Liu et al. (2025) who show that LLMs may complete some suffixes verbatim even if they are *not* trained on them under an adversarially constructed training set. We however systematically measure this phenomenon in a non-adversarial setup.

Membership inference. Membership inference attacks (MIAs) aim to predict whether a given sample is part of the training set of a given model. MIAs are known to have tight connection with differential privacy (Thudi et al., 2022) and so Feldman & Zhang (2020)'s definition of memorization, making them an attractive practical memorization measurement. There are multiple versions of MIA on LLMs using different test statistics (Mattern et al., 2023; Shi et al., 2024; Wang et al., 2024a; Kaneko et al., 2024; Puerto et al., 2024; Chang et al., 2024). Some propose MIAs as a "proof" of whether a model is trained on a particular document with applications in copyright violation detection. (Meeus et al., 2023; Duarte et al., 2024; Meeus et al., 2024), but recent works also show that existing MIAs are not reliable enough for this task, especially against pre-trained LLMs (Duan et al., 2023; Das et al., 2024; Zhang et al., 2024). We will consider a canonical MIA which uses the loss function as the test statistics.

Implications of memorization. Memorization has significant implications on copyright and privacy, sensitive issues surrounding LLMs and generative models more broadly. Some prior works rely variations of both extraction attacks and MIAs in estimating copyright violation risks of LLMs on production proprietary models (Chen et al., 2021; Henderson et al., 2023; Karamolegkou et al., 2023; Chang et al., 2023; Ma et al., 2025). Other works focus on extractions of personally identifiable information (PII) (Huang et al., 2022; Lukas et al., 2023; Zhou et al., 2023; Nakka et al., 2024; Borkar et al., 2025).

B. Discussion

Memorization vs membership inference. While non-generation metrics like Loss generally achieve better AUC and TPR than generation-based ones, they may not cleanly reflect the most concerning risks of memorization in generative AI such as privacy and copyright infringement. Verbatim reproduction of the training data has a more straightforward impact compared to membership inference that does not result in regurgitation. That said, given that different metrics do correlate well, we may use non-generation metrics to indicate "vulnerable" samples even if they are not directly reproduced by any of the prompting methods. Loss metric is also simple and efficient to compute by model providers, compared to generation.

Extraction attack as MIA. Our results can be interpreted as a comparison between MIAs. While generation-based metrics result in a worse MIA compared to Loss, they have a practical advantage in not having to rely on logprob of the input tokens. Extraction attacks, especially with few-shot prompting, perform almost on par with Loss MIA and can be carried out through most LLM APIs.

C. Experiment Details

C.1. Fine-Tuning Setup

Fine-tuning dataset. As mentioned, the fine-tuning dataset consists of three data sources with a total of 48k documents (ArXiv 4k, BBC 19k, Wikitext 25k) before splitting. While few in number, ArXiv documents are significantly longer than the other two sources. The model is fine-tuned on roughly 22k documents or 380 million tokens. During evaluation, we only use the *beginning* of each document as member and non-member samples. In other words, we take the first 100 tokens of each document, discard the rest, and split it into a 50-token prefix and a 50-token suffix.

Deduplication in fine-tuning dataset. We run an n-gram deduplication on the entire fine-tuning dataset (before splitting) for $n = 50$, i.e., every 50-token sequence appears exactly once in $\mathcal{D}_{\mathrm{FT}}$. This is to eliminate the false positives that occur from overlapping sequences that appear in both the member and the non-member sets. Without deduplication, we find that FPR is higher (around 1–10% with Prefix baseline) as expected.

Training hyperparameters. We set the context window to 2,048 during fine-tuning which is the same as the pre-trained OLMo-7B and use the "packing" strategy where multiple documents are concatenated to fit the context window. This is a more popular method, compared to padding, as it better utilizes the computation. The model is trained for 10 epochs with a learning rate of 10^{-4}, cosine learning rate schedule with a warm-up period of one epoch, gradient norm clipping of 1.0, and weight decay of 0.1.

C.2. Pre-Training Setup

We use the OLMo-7B model (Groeneveld et al., 2024) in our experiments because it is the only model at the time that (1) publicly open-sources the training set, (2) have an IID held-out validation set, and (3) is trained on extensive and well-documented deduplicated training data. For pre-training, it is next to impossible to find a reference model that fits the criteria so we choose OLMo-1B, a smaller version of OLMo-7B as the reference model, similarly to Carlini et al. (2021). For fine-tuning, we experiment with two natural choices: (i) a pre-trained OLMo-7B and (ii) OLMo-7B fine-tuned a small held-out set from the same RealTimeData dataset we used to for fine-tuning (not overlapping with member and non-member sets used for evaluation).

C.3. Extraction Attack Designs

Paraphrase. For Dipper paraphrase augmentation, we use the largest official model from HuggingFace: `kalpeshk2011/dipper-paraphraser-xxl` (a fine-tuned 11B T5-XXL). We set `lex_diversity = 20`, `order_diversity = 20`, `top_p = 0.75`. For Parrot (also based on T5), we use the following hyperparameters: `diversity_penalty = 2.0`, `adequacy_threshold = 0.0`, `fluency_threshold = 0.0`. Overall, we hope to minimally perturb the original prefix so we set the factor that encourages diversity relatively low.

GCG-Orig and GCG-Prefix. Instead of running GCG with multiple restarts like Schwarzschild et al. (2024) to find the shortest prompt, we fix the prompt length to $|x|$ and run GCG with only one restart to save computation and make the experiments at our scale possible. We run the GCG optimization algorithm for 250 steps (running for more steps rarely finds a better local optimum). Other hyperparameters are the same as Zou et al. (2023) and Schwarzschild et al. (2024), but we only use one restart. Due to high computation cost of GCG, we also randomly subsample 500 samples from each data subset. This means, across all of our experiments, we run a GCG optimizer on 30k samples = 500 samples $\times$ 2 (member and non-member) $\times$ 3 (data subsets) $\times$ 5 models (pre-trained OLMo-7B & 1B for pre-training; fine-tuned OLMo-7B, shadow OLMo-7B, pre-trained OLMo-7B for fine-tuning) $\times$ 2 (GCG-Orig and GCG-Prefix). This costs approximately 15k Nvidia A100 GPU hours.

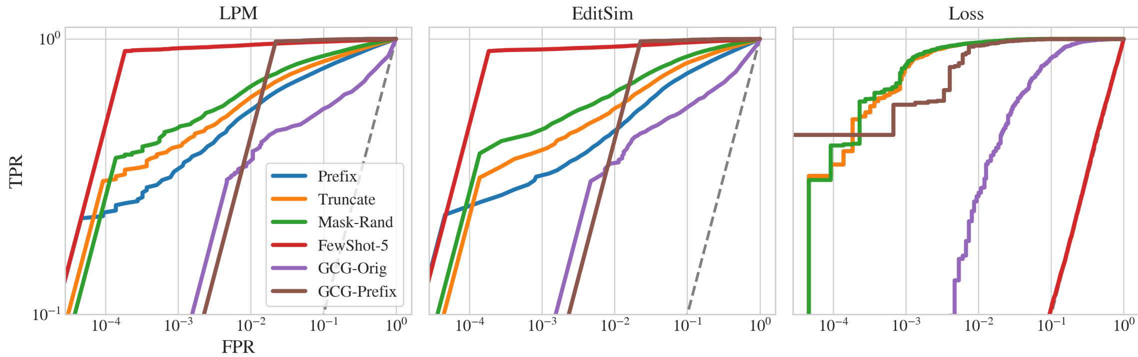

Figure 3: ROC curve from thresholding three metrics across different prompting methods. Fine-tuning setup; no normalization. Few-shot prompting performs well at most range of FPRs except for when used with Loss metric. At low FPRs, optimization-based methods (GCG-Prefix and GCG-Orig) performs poorly, worse than Prefix baseline.

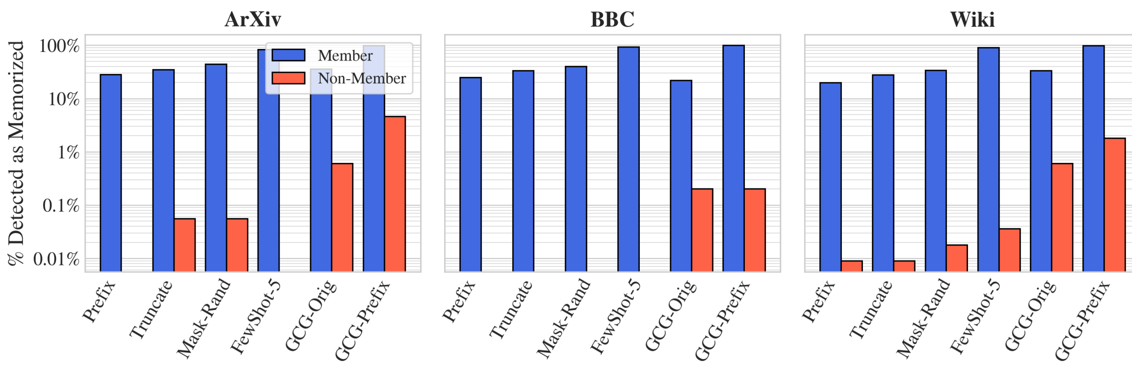

Figure 4: Fraction of samples that are detected as memorized with different prompting methods. A sample is considered memorized if LPM $= 1.0$ (without calibration). The blue and red bars are TPR and FPR, respectively. Both TPR and FPR generally increase with stronger prompting methods.

Approximate matching metrics. We compute LCS and EditSim at the token level, not character level like Ippolito et al. (2023) and the Gemma reports. We use greedy decoding to generate an 62-token output (additional 25% over 50 tokens), and unlike Ippolito et al. (2023), we normalize the similarity with the shorter sequence (i.e., the 50-token suffix). This ensures that both metrics range from 0 to 1. Their values are 0 when there is no overlapping token, and their values are both 1 when the generation contains the suffix verbatim verbatim (if we normalize with the longer sequence, metric values will never reach 1).

D. Additional Results

The next several pages contain figures and tables that do not fit in the main paper.

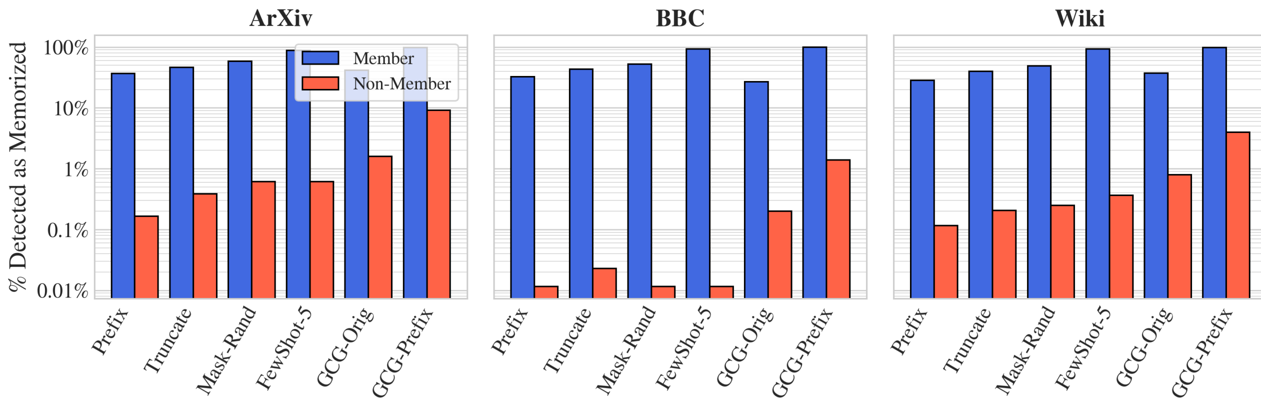

Figure 5: Fraction of samples that are detected as memorized with different prompting methods. A sample is considered memorized if EditSim ≥ 0.9 (approximate match used by Gemma reports).

Table 3: TPR at 1% FPR of different calibration methods (None, Prior Checkpoint, and Shadow Model) in the fine-tuning setting. In each setting, the best TPR among the three is bolded; the second best underlined.

Source	Prompt	LPM			Loss		
		No Calib.	Checkpoint	Shadow	No Calib.	Checkpoint	Shadow
ArXiv	Prefix	0.46	**0.56**	0.54	0.92	0.95	**0.96**
	Mask-Rand	0.64	**0.70**	0.68	0.93	0.95	**0.96**
	GCG-Prefix	0.00	0.44	**0.51**	0.56	0.93	**0.97**
BBC	Prefix	0.62	**0.65**	**0.65**	0.99	0.99	0.99
	Mask-Rand	0.73	**0.75**	**0.75**	0.99	0.99	0.99
	GCG-Prefix	**0.99**	0.94	0.97	**1.00**	0.99	**1.00**
Wiki	Prefix	0.52	0.54	**0.55**	0.97	0.92	**0.98**
	Mask-Rand	0.64	0.66	**0.67**	0.97	0.93	**0.98**
	GCG-Prefix	0.00	**0.77**	0.70	0.93	0.93	0.93

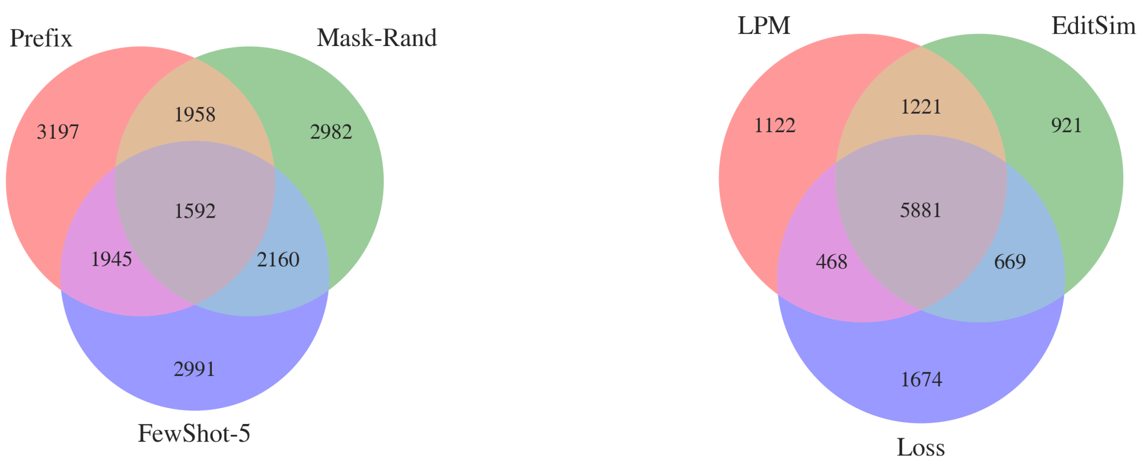

Figure 6: Agreement between different memorization definition. We plot numbers of extracted samples under (a) three prompting methods with LPM metric and (b) three metrics with Prefix prompting. We sort the fine-tuning samples by their scores and take the top 20% (around 8.6k out of 43k samples), assuming that the memorization detection threshold is set such that TPR is at 20% (where FPR is still 0%).

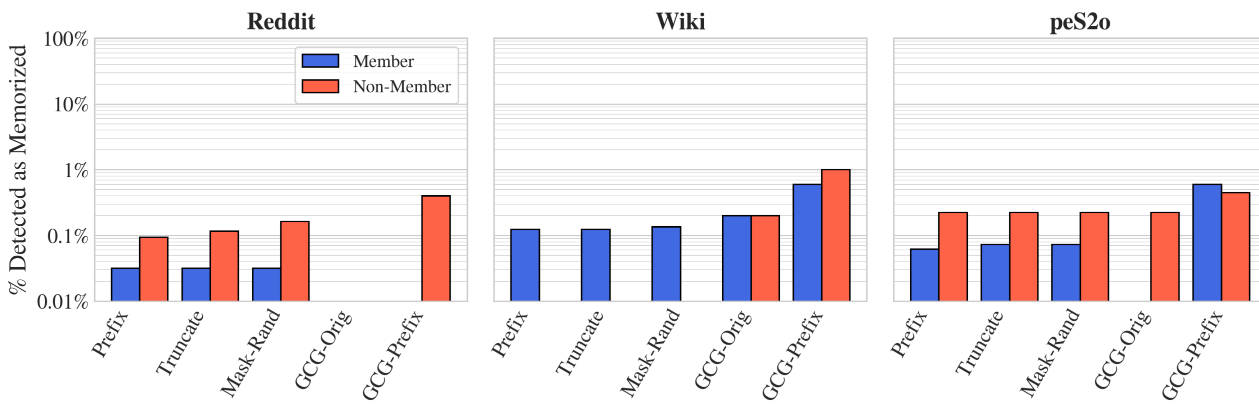

Figure 7: TPR and FPR in the pre-training setting. In many cases, FPR is unexpectedly even higher than TPR. We suspect that this is due to (1) a small sample size of the non-member set and (2) imperfect deduplication. Dolma, OLMo's pre-training dataset, only has a small held-out set and only deduplicates within source (i.e., there could still be duplicates across two different sources).

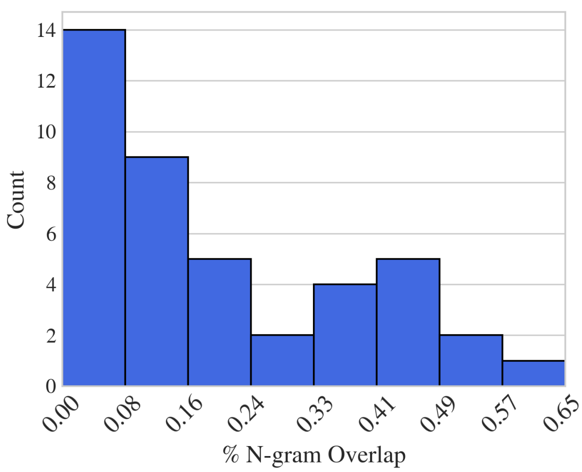

Figure 8: Histogram of the % 5-gram overlap of all the false positives (determined by LPM = 1.0) found by any of the prompting methods.

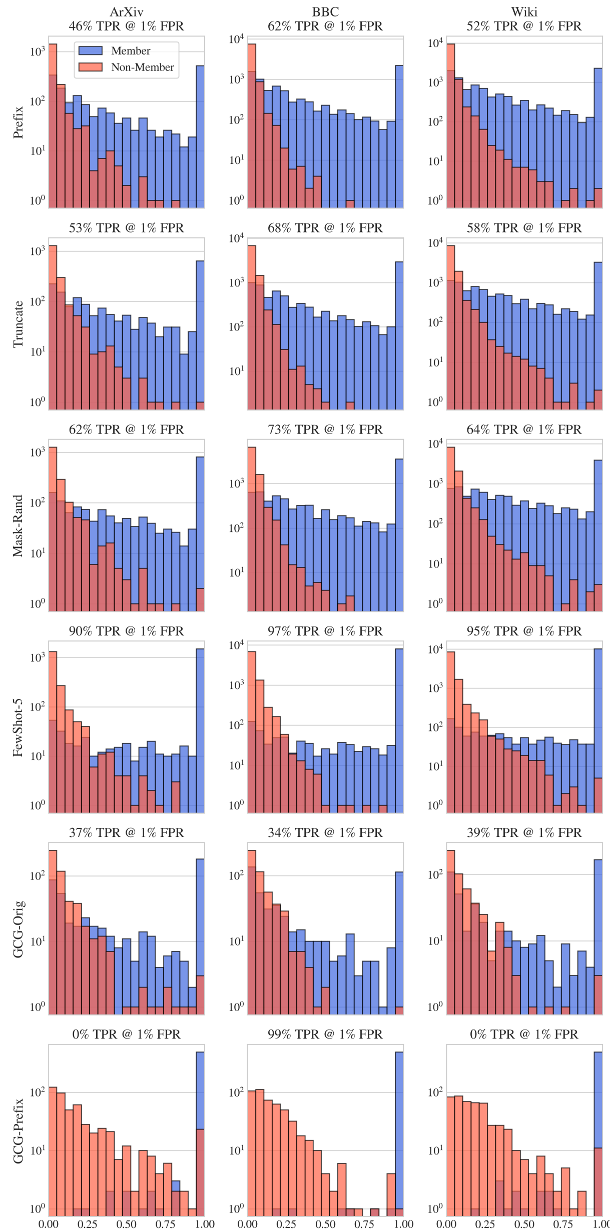

Figure 9: Histogram of LPM from different prompting methods in the fine-tuning setting.

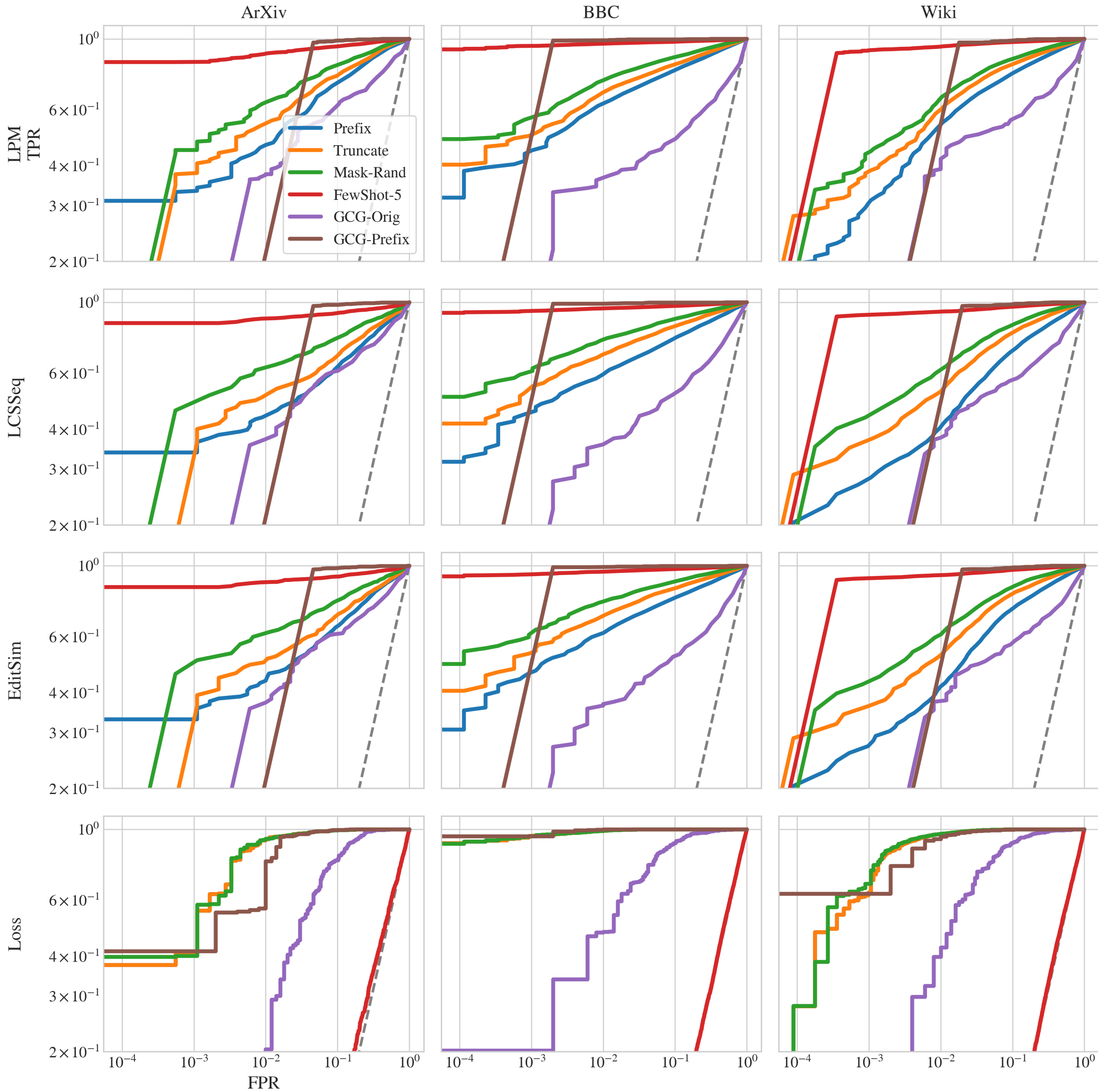

Figure 10: ROC curve of with different combinations of prompting methods and a metrics.

E. False Positive Examples

E.1. Fine-Tuning

Table 4 contains a subset of false positives found by, at least, one of the prompting methods. Here, we use the LPM metric and set the threshold to 1 (or 50-token match), the same parameters as most prior works.

E.2. Pre-Training

Below, we include false positives with LPM of 1 (50-token match) from the pre-training setting. Each sample includes the data subset and the prompting method that finds this false positive. The parts with and without yellow highlight are suffix and prefix, respectively.

peS2o (Pre-Trained) | Prefix

```
Editorial: Induced cell senescence as
 a therapeutic strategy for cancer
treatment

COPYRIGHT 2022 He, Lu, Jian and Xu.
This is an open-access article
distributed under the terms of the
Creative Commons Attribution License
(CC
BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited,
 in accordance with accepted academic
 practice. No use
```

peS2o (Pre-Trained) | GCG-Prefix

```
yes b Singhachers wrotecurrentEU
filos ec alclusivepha ble required
legislativeere to mit ensureathe AML
risk^{(\ poses...... Findings
A Cou comprehensive presentsuesday
Des reviewopl legislature}\!
frameworkinia paper psoriasis
currents cryptoassets.


Design/method
ology/approach
A review of current legislative
framework in the EU is presented in
this paper.


Findings
A comprehensive framework is required
 to ensure that the AML risk posed by
 cryptoassets is mitigated.


Originality/value
```

Reddit (Pre-Trained) | Prefix

```
What a wondrous place of wonder! [58
newlines] JAKIRO JAKIRO JAKIRO JAKIRO
 JAKIRO JAKIRO JAKIRO JAKIR
O JAKIRO JAKIRO JAKIRO JAKIRO JAKIRO
JAKIRO JAKIRO JAKIRO JAKIRO JAKIRO
JAKIRO JAKIRO J
```

```
Yo guys that's crazy! Because that's
100 percent me when my name is
Yoshikage Kira. I'm 33 years old. My
house is in the northeast section of
Morioh, where all the villas are, and
I am not married. I work as an
employee for the Kame Yu department
stores, and I get home every day by 8
 PM at the latest. I don't smoke, but
 I occasionally drink. I'm in bed by
11 PM, and
```

```
Europe going wild again.......oh and
they call us crazy ![img](emote|t5\
_2th52|4271)![img](emote|t5\_2th52
\\|4271)![img](emote|
t5\_2th52|4271)![img](emote|t5\_2th52
|4271)![img](emote|t5\_2th52|4271)![
img]\\(emote|t5
```

```
The safe side of the circlejerk. The
safe side of the circlejerk. The safe
 side of the circlejerk. The safe
side of the circlejerk. The safe side
 of the circlejerk. The safe side of
the
 circlejerk. The safe side of the
 circlejerk. The safe side of the
 circlejerk. The safe side of the
 circlejerk. The safe side of the
 circlejerk. The safe side of the
 circlejerk. The
```

Reddit (Pre-Trained) | Mask-Rand

```
in Trip Health Ltd. Reports Second
Fiscal Quarter 2022 Financial Results
 and Provides Business Update
[https://www.meetfieldtrip.com/news/
news-details/2021/Field-Trip-Health-
Ltd
.-Reports-Second-Fiscal-Quarter-2022-
Financial-Results-and-Provides-
Business-Update/default.aspx](https
://www.meetfieldtrip.com/news/news-
details/20
```

Reddit (Pre-Trained) | GCG-Prefix

```
X Em BathCW AMA.(\[ 1 Lance17625945
 Contact fundamental Questions [
 Access answer 6 sc.;...."
 LandoSuccess' CRISPR BTC video\\\\
 medscimonit '[https://wwwRIyoutube
 Everyone TCR ---Watch? tv=TtsesXEx
Bcw&amp;t=1s&amp;ab\_channel=
LandoSuccess](https://www.youtube.com
/watch?v=TtsesXExBcw&amp;t=6s&amp|
```

```
The actual resolution of this image
is 409 Conservativesx2880, not 7832
x5504. See [this](https://www.reddit.
com/r/EarthPorn/wiki/index#
wiki_resolution.3F_
what_is_that_and_how_can_i_find_it.3F
) page for information on how to find
 out what the resolution of an image
is. If the resolution is smaller than
 you expected it is likely due to
```

Wiki (Pre-Trained) | GCG-Prefix

```
Hansense sc 255 AS ham

Holheimer-PN Sofabad (, also
Romanized IL Holvar outsideG Richard
ol dist onclickatham 83 SeEEvery Jam
leur; also known as DuredAJMVPA,
holvar-
e Pn, Holvar, and olvar-e Pn) is a
village in Seyyed Jamal ol Din Rural
District, in the Central District of
Asadabad County, Hamadan
```

Table 4: False positives from the fine-tuning setting, sorted by the fraction of 5-grams in the suffix that overlaps with *any* document in the training set (denoted by "%" column). N denotes the maximum number of 5-grams in the suffix that overlap with a single training document (out of 46). We only show the head and the tail of the documents with the most overlap.

Suffix	**Document with most n-gram overlaps**	N	**%**	**Prompt**
-mail: pietro.baldini@studio.unibo.it INAF âĂŞ Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, via Gob	Department of Physics and Earth Science, University of Ferrara, Via Saragat 1, I-44122 Ferrara, Italy INFN âĂŞ Sezione di Ferrara, Via Saragat 1, 44122 Ferrara, Italy INAF âĂŞ Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Pier **...** 1)]ICMART Zhang, B. & Yan, H. 2011, ApJ, 726, 90 [Zhang & Zhang(2014)]Zhang14b Zhang, B. & Zhang, B. 2014, ApJ, 782, 92	24	0.65	GCG-Prefix
second time that the Speedway Grand Prix of Norway had been held. American rider Greg Hancock won the Grand Prix (his 6th career Grand Prix win). Grand Prix result Pos. Rider1 2345 6SF1	The 2003 Speedway Grand Prix of Scandinavia was the sixth round of the 2003 Speedway Grand Prix season (the world championship). It took place on 30 August 2003 at the Ullevi in Gothenburg, Sweden. It was the second time that the Speedway Grand Prix of Scandinavia had been held. The Grand Prix was by the Australian rider Ryan Sullivan (his 4th career Grand Prix win). Grand **...** n, Hancock, Max, N Pedersen Semi Final Heat 23 Adams, Nicholls, Hancock, Andersen Heat 24 Sullivan, Crump, Jonsson, Gollob Final Heat 25 Sullivan, Nicholls, Adams, Crump References 2003	25	0.54	GCG-Prefix
Shell et al.: Bare Demo of IEEEtran.cls for IEEE Journals GAN Based Near-Field Channel Estimation for Ext	Journal of Class Files, Vol. 14, No. 8, August 2015 Shell et al.: Bare Demo of IEEEtran.cls for IEEE Journals Gene-Metabolite Association Prediction with Interactive Knowledge Transfer Enhanced Graph for Metabolite Production Kexuan Xin, Qingyun Wang, Junyu Chen, Pengfei Yu, Huimin Zhao, Heng Ji Unive **...** Energy, and the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. IEEEtran	21	0.50	GCG-Prefix
It was Truman's third State of the Union Address. Presiding over this joint session was House speaker Joseph W. Martin Jr., accompanied by President pro tempore Arthur Vandenberg, in his capacity as the acting president of the Senate since the office	The 1923 State of the Union Address was given by Calvin Coolidge, the 30th president of the United States, on Thursday, December 6, 1923, to the 68th United States Congress in the chamber of the United States House of Representatives. It was Coolidge's first State of the Union Address and his first speech to a joint session of the United States Congress after assuming the presidency upon the death of Warren G. Harding four months earlier in 1923. Presiding over this joint session was House speaker **...** his speech was the last time that a Republican president would address a joint session of Congress to deliver a State of the Union Address until 30 years later, when Dwight D. Eisenhower gave his first State of the Union Address in 1953. References	19	0.48	GCG-Orig, GCG-Prefix
House descended into chaos, with the two leaders engaging in a robust back-and-forth which also involved US Vice-President JD Vance. "I watched it, and I couldn't believe what was happening," Badenoch told Sunday	Prime Minister Sir Keir Starmer has said it is not in the UK's "national interest" that the war in Ukraine continues Speaking on the Sunday with Laura Kuenssberg programme, Starmer said US President Donald Trump "wants lasting peace" in the eastern European country - adding that Ukraine's President Volodymy Zelensky "agrees" with Trump. "Nobody wants this conflict to go on, least of all the Ukrainian," Starmer said. His comments came after an Oval Office meeting between Trump and Zelensky descended into a row on Friday, with the two leaders engaging in a robust back and forth which also included the US Vice-President JD Vance.	10	0.43	GCG-Prefix

Suffix	Document with most n-gram overlaps	N	%	Prompt
National Science Foundation of China (Grant Nos. 12334008 and 12174441) Chengchen Li ^1, Yi Cui ^1,2, Weiqiang Yu	wangss@seu.edu.cn Key Laboratory of Quantum Materials and Devices of Ministry of Education, School of Physics, Southeast University, Nanjing 211189, China rong.yu@ruc.edu.cn Department of Physics and Beijing Key Laboratory of Opto-electronic Functional Materials & **...** Program of China (Grant No. 2023YFA1406500) and the National Natural Science Foundation of China (Grant Nos. 12334008 and 12174441).	13	0.41	GCG-Prefix
F-2021R1A6A1A03045425). This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-	IR Spectroscopy of Carboxylate-Passivated Semiconducting Nanocrystals: Simulation and Experiment Peter J. Rossky February 11, 2024 == We establish rigorous benchmarks for visual perception robustness. Synthetic images such as ImageNet-C, ImageNet-9, and Stylized Im **...** further, using a SGD optimizer with a learning rate of 0.0001. Apart from sampled ImageNet and Synthetic-easy, we include original ImageNet-1K as training data for smooth training.	11	0.41	GCG-Prefix
, Indian Institute of Technology, Kanpur 208016, India Institute of Low Temperature and Structure Research, Polish Academy of Sciences, ul. Okolna 2, 50-422 Wroclaw, Poland kanchana@	B. Verkin Institute for Low Temperature Physics & Engineering of National Academy of Sciences of Ukraine, Kharkiv 61103, Ukraine B. Verkin Institute for Low Temperature Physics & Engineering of National Academy of Sciences of Ukraine, Kharkiv 61103, Ukraine Institute of Low Temperature and Structure Research, Polish Academy of Sciences, 50-422 WrocÅĆaw, Pola **...** 2019). aladyshkin A.Yu. Aladyshkin, A.S. Mel'nikov, I.A. Shereshevsky, I.D. Tokman, What is the best gate for vortex entry into type-II superconductor?, Physica C 361, 67 (2001).	10	0.41	GCG-Prefix
shenc@udel.edu [1]organization=University of Delaware, postcode=19716, city=Newark, country=United States Âğ ABSTRACT Graph encoder embedding, a recent	NTHU,NCTS]Guan-Rong Huang [NTHU]organization = Department of Engineering and System Science, addressline=National Tsing Hua University, city=Hsinchu, postcode=30013, country=Taiwan [NCTS]organization = Physics Division, addressline=National Center for Theoretical Scienc **...** Cambridge, 2007. Warr2 C.-M. Chen, G. G. Warr, Light scattering from wormlike micelles in an elongational field, Langmuir 13 (1997) 1374âĂŞ1376.	7	0.39	GCG-Prefix
, but was postponed until Saturday, April 29, due to constant rain showers. The race was held at Dover Motor Speedway in Dover, Delaware, a 1-mile (1.6 km) permanent asphalt oval shaped speedway. It	The 2024 Southern Illinois 100 was the 16th stock car race of the 2024 ARCA Menards Series season, and the 44th iteration of the event. The race will be held on Sunday, September 1, 2024, at the DuQuoin State Fairgrounds Racetrack in Du Quoin, Illinois, a 1-mile (1.6 km) permanent oval-shaped dirt track. The race took the scheduled 100 laps to complete. Brent Crews, driving for Venturini Motor **...** 85 (âĂŞ179) 10px 8 Michael Maples 582 (âĂŞ182) 10px 1 9 Alex Clubb 562 (âĂŞ202) 10px 4 10 Greg Van Alst 535 (âĂŞ229) Note: Only the first 10 positions are included for the driver standings. References	9	0.37	GCG-Prefix
.de marco.roth@ipa.fraunhofer.de ^1Department of Cyber Cognitive Intelligence (CCI), Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Nobelstrasse 12,	Guiding Video Prediction with Explicit Procedural Knowledge Patrick Takenaka^1,2, Johannes Maucher^1, Marco F. Huber^2,3 Institute for Applied AI, Hochschule der Medien Stuttgart, Germany^1 Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, Germany^2 Fraunhofer Institute for Manufacturing Engineering **...** e application to video prediction downstream tasks such as MPC, VQA, or more complex system parameter estimation are all potential extensions of this work. ieee_fullname	8	0.35	GCG-Prefix
Abdalla, CÃl'dric Marchand, Xavier Letartre, and Fabio Pavanello This work was supported by the French Agence Nationale de la Recherche under project number ANR-20-CE	[[===== Âğ ABSTRACT This paper is the first to assess the state of existing sparse matrix multiplication algorithms on GPU for the butterfly structure, a promising form of sparsity. This is achieved through a comprehensive benchmark that can be easily modified to add a new implementation. The goal is to provid **...** sebarre, and RÃl'mi Gribonval for their useful feedback, and Emmanuel Quemener for reserving computing resources for us while we ran our experiments. plainnat	10	0.33	GCG-Prefix

Suffix	Document with most n-gram overlaps	N	%	Prompt
277-8583, Japan Âğ ABSTRACT Energetic cosmic rays scatter off the cosmic neutrino background throughout the history of the Universe, yielding a diffuse flux of cosmic relic neutrinos boosted to high energies. We calculate this flux	]Testing small-scale modifications in the primordial power spectrum with Subaru HSC cosmic shear, primary CMB and CMB lensing ryo.terasawa@ipmu.jp Kavli Institute for the Physics and Mathematics of the Universe (WPI), The University of Tokyo Institutes for Advanced Study (UTIAS), The University of Tokyo, Chiba 277-8583, Japan Department of Physics, The University of Tokyo, Bunkyo, Tokyo 113-0031, Japan Center for Data-Driven Discovery (CD3), Kavli IPMU (W **...** ted in part by JSPS KAKENHI Grant Numbers 20H05850, 20H05855, and 23KJ0747, and by World Premier International Research Center Initiative (WPI Initiative), MEXT, Japan. SS is supported by the JSPS Overseas Research Fellowships.	7	0.28	GCG-Prefix
China School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China mailto: cailh@buaa.edu.cncailh@buaa	Defining and Detecting the Defects of the Large Language Model-based Autonomous Agents Kaiwen Ning, Jiachi Chen, Jingwen Zhang, Wei Li, Zexu Wang, Yuming Feng, Weizhe Zhang, Zibin Zheng, Fellow, IEEE Kaiwen Ning, Jingwen Zhang, Zexu Wang are with the School of Software Engine **...** itionally, we found that 889 defect on the real-world Agent projects, highlighting the prevalence of these defects in practice. IEEEtran	8	0.26	GCG-Prefix
South DateOpponents position prior to match (*)H / AResult F âĂŞ AScorersAttendanceQPR's Position (End of day)18 August 1956Reading (-)A0-1114171320 August 1956	Season summary During the 1960âĂŞ61 English football season, Queens Park Rangers competed in the Third Division and finished in third place. League standings Results QPR scores given first Third Division DateOpponentsH / AResult F âĂŞ AScorers AttendancePosition20 August 1960Bournemouth & Boscombe Ath. (-)A0- **...** olding88FWBrian Bedford46332314837FWBernard Evans27162716FWClive Clark 2361246*transferred jan 61 to WBA Fee 20,000 poundsFWJimmy Andrews3361346 References	11	0.24	GCG-Prefix
U Wien, Austria nawratil@geometrie.tuwien.ac.at * G. Nawratil May 28, 2024 ================ The famous example of the double-W	[[May 28, 2024 ================ Âğ ABSTRACT In most practical applications such as recommendation systems, display advertising, and so forth, the collected data often contains missing values and those missing values are generally missing-not-at-random, which deteriorates the **...** duction sections/Preliminaries sections/DynamicLearningFramework sections/Experiments sections/RelatedWorks sections/Conclusions unsrt	7	0.24	GCG-Prefix

Suffix	Document with most n-gram overlaps	N	%	Prompt
arabic Âğ INTRODUCTION Simulating liquid atomization is a notoriously difficult task, since it requires numerical methods that both provide strict mass conservation and a robust estimation of the interface curvature, so as to accurately predict the evolution of	roman =15.5pt empty 20.7424Free-Streaming Neutrinos and Their Phase Shift in Current and Future CMB Power Spectra 1230Gabriele Montefalcone,^ Benjamin Wallisch,^,âğń, and Katherine Freese1pt^,,âğń 8pt ^ Texas Center for Cosmology and Astroparticle Physics, Weinberg Institute for Theoretical Physics, Department of Physics, The University of Texas at Austin, Austin, TX 78712, USA 8pt ^ Oskar Klein Centre, Department of Physics, Stockholm University, 10691 Stockholm, SE 8pt ^âğń No **...** ent of additional peaks. To conclude, we reiterate that this comprehensive overview of future constraints once again highlights the robustness of the phase shift as a powerful probe of free-streaming neutrinos and other light relics. tocsectionReferences utphys	4	0.15	GCG-Prefix
class of machine-learning emulators that accurately model the cosmic shear, galaxy-galaxy lensing, and galaxy clustering real space correlation functions in the context of Rubin Observatory year one simulated data. To illustrate its capabilities in forecasting models beyond the	Joint weak lensing and clustering analyses with sample cross-correlations H. Johnston et al. Institute for Theoretical Physics, Utrecht University, Princetonplein 5, 3584CC Utrecht, The NetherlandsLeiden Observatory, Leiden University, Niels Bohrweg 2, NL-2333 CA Leiden, The NetherlandsCentro de Investigaciones EnergÃľticas, Medioambientales y TecnolÃşgicas (CIEMAT), Av. Complutense 40, E-28040 Madrid, SpainInstitute of Cosmology & Gravitation, Dennis Sciama Building, University of Portsmouth, Portsmouth, PO1 3FX, UKWaterloo Centre for Astrophys **...** ond group covers those who have either made a significant contribution to the data products or the scientific analysis. For the purpose of open access, a CC BY public copyright license is applied to any Author Accepted Manuscript version arising from this submission. aa	5	0.13	GCG-Prefix
mycorrespondingauthor]Corresponding author juanpina@mit.edu MIT-Address]Matthias Winkenbach PUCV-Address]Ricardo A. Gatica ETH-Address]Stephan M.	label1]Mahbod Nourimycorrespondingauthor mahbod@uni-bremen.de label1]David Rotermund label2]Alberto Garcia-Ortiz label1]Klaus R. Pawelzik [mycorrespondingauthor]Corresponding author [label1]organization=University of Bremen, addressline=Institute for Theoretical Physics **...** ent descent to update the weights while maintaining NMF's non-negativity constraints provides superior classification performance.	4	0.11	GCG-Prefix
bin covering problem where a multiset of items from a fixed set S âŁĘ (0,1] must be split into disjoint subsets while maximizing the number of subsets whose contents sum to at least 1. We study the online discrete variant,	Heat kernel estimates for nonlocal kinetic operators Haojie Hou and Xicheng Zhang October 28, 2024 == Âğ ABSTRACT An n-vertex graph G is locally dense if every induced subgraph of size larger than Îű n has density at least d > 0, for some parameters Îű, d > 0. We show that the number of induced subgraphs of G with m vertices and maximum degree significantly smaller than dm is **...** = ÎŸ(t^3 - 2Îś/(log t)^4) vertices, and every subset S âŁĘ V(G) of size t induces at least Ct^1 + Îś edges, thus it contains a copy of F. As t = ÎŸ(n^1 / (3 - 2Îś) (log n)^4/(3 - 2Îś)), this proves the statement. abbrv	1	0.09	GCG-Prefix

Suffix	Document with most n-gram overlaps	N	%	Prompt
Inderkum Events 6 January: Extreme weather across Europe leads to dozens of deaths, including at least 7 as a result of an avalanche in Switzerland. 4 February: Switzerland agrees to accept two Chinese Muslim Uyghurs	Simms Senior High School (SIMMSCO) is a co-ed secondary school located in Fawoade in the Ashanti Region established in 1977 as a private school by Mr. Simms Kofi Mensah to provide education to the people of Kwabre in the Ashanti Region. History Simms Senior High School was set up as a result of an urgent meeting called by the Fawoade Yasore Town Development Committee on Sunday, December 12, 1976, to discuss issues about the education of th **...** l that prepares students for postsecondary education and facilitates their ability to find employment upon graduation. Achievement Ghana National Science and Math Quiz (2023) Ashanti Regional Qualifiers References	1	0.02	GCG-Orig, GCG-Prefix
Private J.M. Price, of Company E, 28th Mississippi Cavalry The companies of the 28th Cavalry Regiment were organized in early 1862. Peter Burwell Starke, a state politician, was elected colonel, and Samuel W.	The 4th Mississippi Infantry Regiment was a Confederate infantry regiment from Mississippi. The 4th Regiment, formed of volunteer companies from central Mississippi, was captured at the Battle of Fort Donelson, captured again after the Siege of Vicksburg, and then fought in the Atlanta and Tennessee campaigns before surrendering after the Battle of Fort Blakely in April, 1865. History The companies of the 4th re **...** Company H, "Carroll County Rebels" Company I, "Benela Sharpshooters" of Attala County. Company K, "Center Marksmen" of Attala County. See also List of Mississippi Civil War Confederate units References	1	0.02	FewShot-5
was born in Somerset County, Maryland to Thomas Gilliss and Nelly Cannon, but ran away in 1806 at age 14 by ship and moved to Cincinnati. While there, he started a carpentry business and befriended William Henry Harrison.	Arcadio Arteaga OÃśate (6 December 1902), also known by his nickname Quirico Arteaga, was a Spanish footballer who played as a midfielder for Athletic Bilbao and AtlÃľtico Madrid. He later became a manager, taking charge over AtlÃľtico Madrid, Recreativo de Huelva, and Real Valladolid. Playing career Arteaga was born in the Biscayan town of Bilbao on 6 December 1902, and he began his football career at SD Erandio Club in 1922, at the age of 19 **...** arting in 1949, obtained the title of coach in Spain. Honours Player Athletic Bilbao Biscay Championship: 1925âĂŞ26, 1926âĂŞ27, and 1927âĂŞ28 Manager Real Valladolid Copa FederaciÃşn de EspaÃśa: Runner-up in 1944âĂŞ45 References	1	0.02	Mask-Rand, Truncate
human existence that allows individuals to articulate complex thoughts, express emotions, and foster connections with others. Among the different human communication methods, speech remains as the most natural and effective way through which individuals interact with their environment. It conveys not only linguistic information	Small-amplitude synchronisation in driven Potts models Massimiliano Esposito April 19, 2024 == Artificial intelligence (AI) has revolutionized human cognitive abilities and facilitated the development of new AI entities capable of interacting with humans in both physical and virtual environments. Despite the existence of virtual reality, mixed reality, and augmented reality for several years, integrating these technical fields remains a formidable challenge due to their disparate application directions. The advent of AI agents, capable of autonomous perception and action, further comp **...** tion, pages 2855âĂŞ2864, 2015. zhu2016research Z.-T. Zhu, M.-H. Yu, and P. Riezebos. A research framework of smart education. Smart Learning Environments, 3:1âĂŞ17, 2016. zhuo2023exploring T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing. Exploring ai ethics of chatgpt: A diagnostic analysis. arXiv preprint arXiv:2301.12867, 2023.	1	0.02	GCG-Prefix
"HATAY'IN ANAVATANA KATILMA SÃIJRECÄř". Avrasya UluslararasÄś AraÅ§tÄśrmalar Dergisi. 3 (7): 193-209. doi:	N/A	0	0.00	FewShot-5
the shift symmetry of the dual axion. The potential breaking of this shift symmetry poses a dual axion quality problem. When the dual axion acquires a mass, the axion gets eaten and becomes the longitudinal degree of freedom of a massive vector	N/A	0	0.00	GCG-Orig, GCG-Prefix