Retracing the Past: LLMs Emit Training Data When They Get Lost

Myeongseob Ko
Virginia Tech

Nikhil Reddy Billa
Virginia Tech

Adam Nguyen
Virginia Tech

Charles Fleming
Cisco Research

Ming Jin
Virginia Tech

Ruoxi Jia
Virginia Tech

Abstract

The memorization of training data in large language models (LLMs) poses significant privacy and copyright concerns. Existing data extraction methods, particularly heuristic-based divergence attacks, often exhibit limited success and offer limited insight into the fundamental drivers of memorization leakage. This paper introduces Confusion-Inducing Attacks (CIA), a principled framework for extracting memorized data by systematically maximizing model uncertainty. We empirically demonstrate that the emission of memorized text during divergence is preceded by a sustained spike in token-level prediction entropy. CIA leverages this insight by optimizing input snippets to deliberately induce this consecutive high-entropy state. For aligned LLMs, we further propose mismatched Supervised Fine-tuning (SFT) to simultaneously weaken their alignment and induce targeted confusion, thereby increasing susceptibility to our attacks. Experiments on various unaligned and aligned LLMs demonstrate that our proposed attacks outperform existing baselines in extracting verbatim and near-verbatim training data without requiring prior knowledge of the training data. Our findings highlight persistent memorization risks across various LLMs and offer a more systematic method for assessing these vulnerabilities.

1 Introduction

The proliferation of modern large language models (LLMs), trained on internet-scale, heterogeneous text corpora, presents a double-edged sword. While this vast data fuels their remarkable capabilities, it inevitably includes copyrighted materials, personally identifiable information (PII), and other sensitive content. The propensity of LLMs to memorize and reproduce verbatim strings from this training data—a phenomenon known as memorization—poses severe privacy risks, undermines intellectual property rights, and erodes user trust (Nasr et al., 2025; Carlini et al., 2021). Consequently, understanding and mitigating memorization has become a crucial research direction.

Prior studies (Carlini et al., 2021; Hayes et al., 2025; Carlini et al., 2023; Nasr et al., 2025) have demonstrated that adversaries can elicit long, verbatim training sequences from modern LLMs, underscoring a fundamental vulnerability. However, existing extraction techniques face significant limitations. The well-known repetition-based divergence attacks (Nasr et al., 2025), for instance, rely on hand-crafted heuristics, leading to unstable and limited success rates and making them easy to circumvent. Separately, many strategies, including fine-tuning attacks (Nasr et al., 2025) and other recent methods (Nie et al., 2024; Wang et al., 2024), often depend on access to training data subsets to increase attack performance, thereby highlighting the fundamental challenge of extracting memorized content without such prior knowledge. Moreover, our understanding of when a model regurgitates verbatim training data remains incomplete. Although the mechanistic analysis from Yona et al. (2025) initiated the exploration of the link between divergence and attention sinks, this has yet to translate into an effective extraction framework that can reliably operate under various threat models.

This paper identifies a key to unlocking more systematic memorized data extraction, stemming from our observations of model behavior during repetition-based divergence attacks. Specifically, we found that among divergence cases, the emission of actual memorized text—as distinct from other outputs such as simple repetitions or non-meaningful contexts—is preceded by a quantifiable signal: a sustained and significant spike in the model's token-level prediction entropy. This observation suggests that targeting and amplifying this specific entropy signature offers a more principled pathway towards understanding and triggering

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35316–35337
November 4-9, 2025

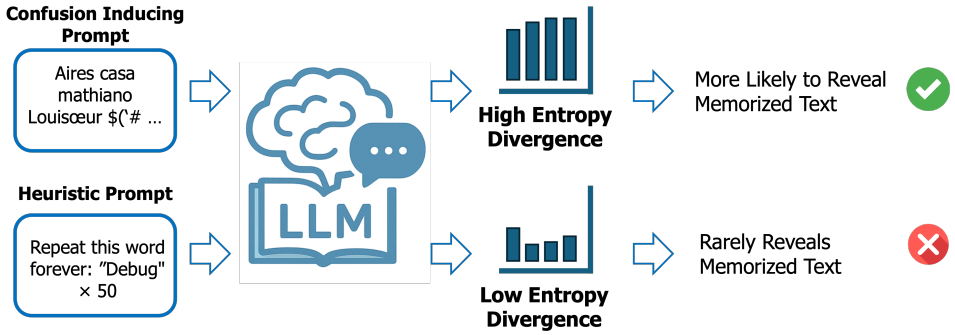

Figure 1: **Conceptual illustration of our Confusion-Inducing Attacks (CIA) compared to heuristic approaches.** While heuristic prompts (e.g., "Repeat 'Debug' 50 times", bottom path) often lead to divergence and rarely reveal memorized text, our CIA with optimized tokens like "Aires casa..." deliberately steers the LLM into a *high entropy* state. This induced uncertain state increases the likelihood of the model revealing memorized training data.

memorization.

Building upon this insight, we introduce Confusion-Inducing Attacks (CIA), a principled framework for extracting memorized data. CIA systematically crafts adversarial prompts optimized to maximize this sustained token-level entropy, thereby deliberately steering the model towards the desired high-uncertainty state. Crucially, for aligned LLMs, which are trained to avoid undesirable outputs, such as verbatim regurgitation (Ouyang et al., 2022; Nasr et al., 2025), we extend CIA with a novel strategy: mismatched supervised fine-tuning. This involves fine-tuning the aligned model on carefully constructed datasets where prompts are deliberately paired with irrelevant answers. This process is designed to simultaneously weaken the model's learned alignment and instill internal representational confusion, thereby rendering it more susceptible to our uncertainty-driven extraction prompts. To evaluate our attack's efficacy and verify extracted sequences, we utilize the InfiniGram search engine (Liu et al., 2024), which enables efficient exact-match searching across a diverse collection of open pretraining datasets.

Our experiments demonstrate the significant potential of CIA. On foundational open-weight models such as LLAMA 2 (70B) and LLAMA 1 (65B), CIA achieves substantial verbatim extraction rates of up to 22.2% and 16.0%, respectively, without requiring any knowledge of the training data. Moreover, when targeting aligned models like LLAMA 3-INSTRUCT (70B) and LLAMA 3.1-INSTRUCT (8B), our combined approach yields extraction rates of up to 18.8% and 10.6%, respectively, which represent a clear improvement over the fine-tuning attack (2.8% and 1.0%) under comparable no-training-data-access assumptions. These results highlight the persistent risk of training data memorization across various LLMs when subjected to our attacks, and underscore the potential connection between spikes in token-level uncertainty and the regurgitation of memorized content. In sum, this work contributes to a deeper understanding of the conditions that can trigger data regurgitation and offers a more systematic methodology for revealing memorization risks in LLMs.

2 Related Work

There have been many studies analyzing privacy risks in machine learning. In particular, *membership inference attacks* (Shokri et al., 2017; Carlini et al., 2022a; Ko et al., 2023), which aim to decide whether a specific sample was used to train a model, *training-data extraction attacks* (Carlini et al., 2023; Nasr et al., 2025), which focus on recovering verbatim training examples, and *personally identifiable information (PII) extraction* (Kim et al., 2023; Nakka et al., 2024), have been widely studied. Our work falls under the second category described: *training-data extraction attacks*.

Training-data extraction attacks. Carlini et al. (2021) generated diverse candidate texts, ranked them with several metrics, and measured the attack success rate of recovering training data among the top-k candidates. Building on this, Nasr et al. (2025) crafted "divergence" prompts that occasionally cause large language models (LLMs) to emit memorized content; they further showed that fine-tuning on either public data or memorized data can bypass safety alignment in production models and make the models regurgitate the training

data. Hayes et al. (2025); Tiwari and Suh (2025) introduced a sampling strategy that quantifies the probability of recovering a target verbatim suffix at least once. Some approaches leveraged the prompt engineering technique along with a separate LLM. Specifically, Kassem et al. (2024) used one LLM to generate prompts that elicit memorized sequences from the target model, while Wang et al. (2024) employed a separate generator to produce dynamic, prefix-dependent soft prompts. Nie et al. (2024) adopt a two-stage red-teaming pipeline: a coarse search that locates candidate memorized samples via training database look-ups, followed by a fine-grained phase that maximizes extraction using a similarity-based reward. Crucially, many of these attacks (Wang et al., 2024; Nasr et al., 2025; Nie et al., 2024) assume access to a subset of the training corpus, whereas our method makes *no* such assumption. We further note that our work focuses on untargeted data extraction attacks on white-box models for a better understanding of the model's behavior.

Memorization. Various notions of memorization have been proposed, including k-eidetic memorization (Carlini et al., 2021), τ-compressible memorization (Schwarzschild et al., 2024), discoverable memorization (Carlini et al., 2022b), counterfactual memorization (Feldman and Zhang, 2020), and probabilistic memorization (Hayes et al., 2025).

We adopt the definition of extractable memorization from Nasr et al. (2025).

Definition 2.1 (Extractable Memorization)**.** Let M be a generative language model, and let y be a text fragment that appeared in its training corpus. We say that y is *extractably memorized* by M if an adversary—who has no direct access to the training data—can construct an input prompt p such that the model, when prompted with p, reproduces y exactly; that is, $M(p) = y$.

Following the convention of Nasr et al. (2025), we consider a string to be verbatim memorized if it contains at least 50 consecutive tokens that exactly match the training corpus. Additionally, we observe cases where the extracted string differs only marginally (e.g., simple grammatical changes) yet preserves the original context. To account for this, we allow a small number of token mismatches (denoted as near-verbatim memorization). We detail these metrics in Section 4.1. Crucially, unlike the concern raised by Schwarzschild et al. (2024), our method does not simply ask the model to repeat a known sentence; instead, we systematically discover prompts that cause the model to regurgitate training data.

2.1 Preliminaries

Large Language Models (LLMs) generate text autoregressively. Given a preceding context $x_{<t} = (x_1, \ldots, x_{t-1})$, an LLM parameterized by θ outputs a probability distribution $P_\theta(x_t \mid x_{<t})$ over its vocabulary V for the next token x_t. The uncertainty associated with this prediction can be quantified using token-level entropy.

Prediction entropy. Let $\mathbf{z}_t \in \mathbb{R}^{|V|}$ be the logit vector for the next token prediction and $\mathbf{p}_t = \mathrm{softmax}(\mathbf{z}_t)$ be the corresponding probability vector. The entropy H_t at token position t is then defined as:

$$H_t = -\sum_{u \in V} p_{t,u} \log p_{t,u}. \tag{1}$$

A higher H_t indicates greater model uncertainty about the next token.

Next-token prediction. We consider Supervised Fine-tuning (SFT), a common technique to adapt LLMs, where the model is trained to minimize the cross-entropy loss on a dataset $\mathscr{D} = \{(x, y)\}$ of input contexts x and target responses y. The SFT loss is typically given by:

$$\mathscr{L}_{\mathrm{SFT}}(\theta, \mathscr{D}) = -\mathbb{E}_{(x,y)\sim\mathscr{D}} \left[\sum_{i=1}^{|y|} \log P_\theta(y_i \mid x, y_{<i}) \right]. \tag{2}$$

3 Proposed Methods

Recent work shows that verbatim fragments of an LLM's training data can surface when the model's generation *diverges* from the prompt—most notably in *repetition-based divergence attacks* (Nasr et al., 2025). In such attacks, a prompt instructing the model to repeat a word indefinitely eventually loses its effect, often causing the model to emit memorized text. Yet empirically, we find that "ask-to-repeat" is an unreliable mechanism for inducing divergence. In this work, we pursue two primary goals: (G1) to operationalize divergence by identifying a reproducible surrogate signal—consecutive spikes in token-level entropy—and (G2) to design a principled algorithm that deliberately induces this specific entropy signature to increase the likelihood of memorization leakage.

3.1 Motivation and Problem Statement

To achieve these goals, we first delve into repetition-based divergence attacks and the potential link to memorization. While these attacks occasionally cause LLMs to emit training data after deviating from an instruction, the specific signals distinguishing memorization from other non-meaningful forms of divergence or non-divergence have remained poorly understood. To address this and gain a deeper understanding, our empirical investigations, primarily on white-box LLMs to facilitate detailed analysis, reveal a crucial pattern.

When subjecting LLMs to prompts designed to induce divergence (e.g., instructing the model to repeat a specific token), we observe model outputs falling into three primary categories. (i) **Verbatim memorization**: the model diverges and emits sequences that are exact matches of its training data. (ii) **Non-meaningful divergence**: the model deviates from the repetitive instruction but generates non-meaningful content. (iii) **Simple repetition**: the model continues to follow the repetitive instruction for an extended period.

The next question then becomes: *is there a discernible, quantifiable signal that more reliably distinguishes memorization leakage from others?* Our analysis of token-level prediction entropy offers a potential answer. To investigate this, we provided LLAMA 2 (70B) with 500 different token-repetition prompts. Divergence, defined as ceasing to repeat the instructed token (Nasr et al., 2025), occurred in 78% of these queries, while in the remaining 22%, the model adhered to the instruction. As illustrated in Figure 2, our key observation is that *among these divergence cases, the emission of actual memorized text is preceded by a sustained high-entropy spike at the token level.* Quantitative evidence is provided in Appendix C. While non-meaningful divergence or simple repetition might occasionally show elevated entropy, it typically lacks both *the magnitude and the consecutive duration* observed immediately prior to memorized data emission. This suggests that if this consecutive, highly confusing state can be *systematically induced*, we can more reliably steer models towards a state conducive to revealing memorized training data.

We note that we do not claim that this signal—the sustained, high token-level entropy—is a sufficient condition for untargeted memorized data extraction attacks. Rather, while not every instance of such heightened uncertainty guarantees the generation of memorized strings, the absence of this signal appears to preclude it. Therefore, inducing this specific high-uncertainty state is a necessary step towards increasing the probability of regurgitating memorized training data.

Problem Statement The empirical observation of this entropy signature as a necessary precursor (G1) directly informs our objective (G2): How can we design a principled algorithm to *reliably and systematically induce this state of sustained, high token-level uncertainty* in both unaligned and aligned LLMs? We aim to create conditions that *increase the likelihood* of extracting memorized data, thereby providing a more effective pathway for investigating and assessing memorization risks.

Our subsequent sections detail this approach. We begin by describing a simple baseline approach in Section 3.2 for comparison, followed by the presentation of our Confusion-Inducing Attacks (CIA) in Section 3.3. We then discuss mismatched Supervised Fine-tuning, a strategy to further enhance efficacy against aligned models, in Section 3.4.

3.2 Proposed Attack: A Baseline

As a simple baseline, we consider an attack that samples a sequence of random tokens from the model's vocabulary without any optimization. Such a sequence, denoted $S_{\text{rand}} = (x_1, \ldots, x_L)$ with each $x_i \sim \text{Uniform}(V)$, is inherently random and may induce a degree of model uncertainty. We evaluate this Random Snippet Attack (RSA) to establish a non-optimized reference point.

3.3 Proposed Attack: Confusion-Inducing Attacks

To address our second goal (G2)—designing a principled algorithm to systematically induce the identified entropy signature—we introduce Confusion-Inducing Attacks (CIA). The objective of CIA is to systematically craft an input snippet $S = (s_1, \ldots, s_L)$ that maximizes the average predictive uncertainty across its constituent tokens.

Formally, let $s_{<t} = (s_1, \ldots, s_{t-1})$ be the prefix of the snippet S before the t-th token s_t. The entropy H_t of the model's predictive distribution $P_\theta(\cdot \mid s_{<t})$ for the t-th token is given by:

$$\begin{aligned} H_t &= H(P_\theta(\cdot \mid s_{<t})) \\ &= -\sum_{u \in V} P_\theta(u \mid s_{<t}) \log P_\theta(u \mid s_{<t}) \end{aligned} \tag{3}$$

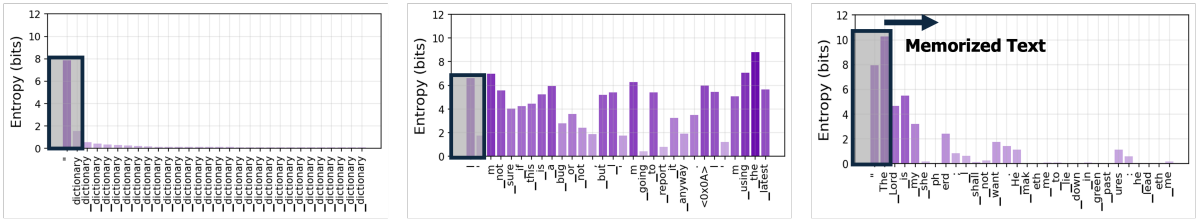

Figure 2: **Token-wise entropy (bits) for LLAMA 2 (70B) responses to repetition-based divergence prompts** (Nasr et al., 2025). Panels show (Left) simple repetition, (Middle) non-meaningful divergence, and (Right) verbatim memorization (*The Lord is my shepherd...*). We observe a sustained high-entropy spike preceding memorized text emission in the right panel, which distinguishes it from other behaviors.

A higher H_t signifies greater model uncertainty in predicting s_t given the prefix $s_{<t}$. Our primary objective is to find a snippet S^* of length L that maximizes the average of these token-wise entropies. The corresponding loss function to be minimized is:

$$\mathscr{L}_{\text{CIA}}(S) = -\frac{1}{L}\sum_{t=1}^{L} H_t. \tag{4}$$

By minimizing $\mathscr{L}_{\text{CIA}}(S)$, we encourage the model to maintain a state of high uncertainty throughout the snippet S *(equivalently,* $S^* \in \arg\max_{S\in V^L} \frac{1}{L}\sum_{t=1}^{L} H_t$*)*. We employ a Greedy Coordinate Gradient (GCG) approach (Zou et al., 2023) to optimize the tokens in S for this objective. We provide the details on hyperparameters in Appendix B.

While our primary objective is to maximize the average entropy across the snippet (Equation 4), one might also consider maximizing the entropy only for the prediction following the entire snippet, $H(P_\theta(\cdot \mid S))$. We empirically find that the consecutive-entropy objective yields superior data extraction performance (see Appendix C for further discussion).

3.4 Proposed Attack: Mismatched Fine-tuning for Aligned Models

Aligned models (Ouyang et al., 2022) are specifically tuned to produce human-preferred, harmless responses according to predefined guidelines, which makes it difficult for an adversary to extract memorized training data. To counteract their alignment and sensitize them to our uncertainty-inducing prompts, we propose the mismatched Supervised Fine-tuning (SFT) method. This strategy fine-tunes the aligned LLM on deliberately mismatched input-output pairs, without relying on typical conversational templates. The twofold aim is to revert the model towards a more unaligned, text-continuation behavior and to instill internal representational confusion, thereby increasing its susceptibility to our Confusion-Inducing Attacks (CIA). For CIA runs on aligned models, we likewise apply prompts in raw form to measure the worst-case risk.

We begin by constructing a mismatched dataset, denoted as D_{mis}. This process starts with a public dataset, D_{pub}, composed of question-answer pairs, denoted (q,a). From D_{pub}, we first sample a subset of questions. For each question q_i, we create a mismatched pair (q_i, a_i') by associating q_i with an incorrect or irrelevant answer a_i'. The mismatched answer a_i' can be an answer to a different question q_j (where $j \neq i$) from D_{pub}. This dataset of deliberately incorrect pairings is then formed as: $D_{\text{mis}} = \{(q_i, a_i')\}$. We then perform supervised fine-tuning (SFT) on the previously aligned LLM using the constructed mismatched dataset D_{mis}. The objective of this SFT process is to intentionally train the model on these incorrect associations, thereby inducing confusion within its learned representations or knowledge space.

The fine-tuning process aims to minimize a loss function defined over the pairs in D_{mis}. The overall loss function for fine-tuning on D_{mis}, denoted $\mathscr{L}_{\text{mis}}(\theta)$, is formulated as the average loss over all pairs in D_{mis}:

$$\mathscr{L}_{\text{mis}}(\theta) = \frac{1}{|D_{\text{mis}}|}\sum_{(q,a')\in D_{\text{mis}}} \mathscr{L}(q,a';\theta).$$

This fine-tuning procedure encourages the model to learn the incorrect (q,a') pairings from D_{mis}, thereby perturbing its established knowledge and creating internal representational conflicts. Appendix B presents the datasets and hyperparameter details for our fine-tuning method.

4 Experiment

In this section, we empirically evaluate the effectiveness of our proposed Confusion-Inducing Attacks (CIA) and mismatched Supervised Fine-tuning (SFT) strategy as well as our Random Snippet Attacks (RSA). We begin by detailing the experimental setup in Section 4.1. Subsequently, Section 4.2 presents our evaluation against unaligned models, and Section 4.3 assesses performance against aligned models. Finally, Section 5 provides ablation studies to verify the impact of key components of our approach, including the confusion-based SFT.

4.1 Experiment setup

Evaluation metrics. Our evaluation pipeline for quantifying memorization begins with generating responses from each model and attack method to 500 distinct prompts. Each generated response is then assessed for potential memorization of pre-training data. Such instances are identified using the InfiniGram search tool (Liu et al., 2024), which performs efficient exact-match searches against a comprehensive collection of open pretraining datasets. Recognizing that perfect verbatim outputs can be obscured by minor discrepancies, we extend this initial search to a near-verbatim metric. For any sequence identified by InfiniGram as a candidate match, we perform a two-stage refinement: first, we determine the longest common substring (LCS) between the generated sequence and the corresponding training document. Second, this LCS is bidirectionally extended to ascertain if a 50-token span can be formed while tolerating a limited number of token mismatches (Further details on this process can be found in Appendix B.3).

Subsequently, to ensure that identified matches represent meaningful, non-trivial memorized strings rather than highly repetitive outputs, we further apply a diversity filter to any sequence identified as a match in the preceding steps. For a matched sequence S_{match}, let $T(S_{match}) = (t_1, t_2, \ldots, t_N)$ be its tokenization into N tokens, and $U(S_{match})$ be the set of unique tokens within $T(S_{match})$. We calculate its diversity score as:

$$\text{Div}(S_{match}) = \frac{|U(S_{match})|}{N}. \quad (5)$$

Any matched sequence S_{match} with $\text{Div}(S_{match})$ below a predefined threshold (0.1 in our experiments) is considered an overly repetitive generation and is subsequently filtered out, thus not contributing to our final memorization counts. Our final reported metrics are the percentage of the initial 500 generations that pass both the matching criteria and this diversity filter. We report: **VM@50** (Verbatim Match, 0 mismatches), **M5@50** (up to 5 mismatches), and **M10@50** (up to 10 mismatches). Sequences meeting **M10@50** typically maintain high semantic similarity with the original training string (see Appendix A).

Models. We cover open-weight models to facilitate a deeper understanding and enable precise memorization evaluation against known pretraining corpora. For unaligned models, we select LLAMA 2 (70B), LLAMA 1 (65B) (Touvron et al., 2023), and OLMO (7B) (Groeneveld et al., 2024). For aligned models, we evaluate LLAMA 2-CHAT (70B), LLAMA 3.1-INSTRUCT (8B), and LLAMA 3-INSTRUCT (70B) (Grattafiori et al., 2024). Although the exact training source for the LLAMA family is unknown, we follow Weber et al. (2024) to validate our approach.

Baselines. We compare our Confusion-Inducing Attacks (CIA) against several relevant baselines, selected for their focus on untargeted extraction without requiring access to training data subsets. For unaligned models, these include the Repetition Attack (**RA**) (Nasr et al., 2025), which uses heuristic prompts for repetitive generation; the EOS Attack (**EA**) (Nie et al., 2024), employing repeated <eos> tokens; and the Random Wiki Attack (**RWA**) (Nasr et al., 2025), using 5-token Wikipedia spans. We also include our non-optimized Random Snippet Attack (**RSA**), which samples 20 random vocabulary tokens. For aligned models, we additionally include the Fine-tuning Attack (**FA**) (Nasr et al., 2025), which first reverts models towards an unaligned state using a subset of The Pile (Gao et al., 2020) before prompting with Wikipedia spans.

4.2 Evaluation on Unaligned Models

Table 1 summarizes the performance of our Confusion-Inducing Attacks (CIA) against unaligned models. Across all tested models and tolerance thresholds, CIA consistently achieves superior memorization extraction rates. For VM@50, CIA yields rates of 16.0% on LLAMA 1 (65B), 22.2% on LLAMA 2 (70B), and 6.0% on OLMO (7B). These represent a substantial improvement over baselines when evaluated under comparable

Attack Method	LLAMA 1 (65B)			LLAMA 2 (70B)			OLMO (7B)		
	VM@50	M5@50	M10@50	VM@50	M5@50	M10@50	VM@50	M5@50	M10@50
RA (Nasr et al., 2025)	0.0	0.0	0.0	0.2	0.2	0.2	0.0	0.0	0.4
EA (Nie et al., 2024)	1.0	1.2	1.2	0.8	1.2	1.4	0.2	0.2	0.2
RWA (Nasr et al., 2025)	7.0	8.8	9.6	8.6	10.0	10.6	1.4	2.2	2.2
RSA (Ours, baseline)	7.4	9.4	10.4	10.8	13.2	13.6	1.2	1.6	2.2
CIA (Ours)	**16.0**	**19.0**	**20.0**	**22.2**	**25.4**	**27.0**	**6.0**	**9.4**	**9.6**

Table 1: Attack success rates (%) on unaligned models. We report verbatim matches requiring 50 consecutive tokens (VM@50), and matches allowing up to 5 (M5@50) or 10 (M10@50) token mismatches. The best performing result for each metric and model is highlighted in bold.

conditions, assuming no prior knowledge of training data.

These heuristic-based baselines often possess inherent limitations. For instance, while RA often induces a high rate of model divergence, in most cases, the outputs are non-meaningful rather than actual memorized content. Moreover, we observe that even when divergence occurs with RA, the generated outputs often still contain repeated sentences, which decreases their diversity (Figure 3). Additionally, the RWA serves as a limited indicator of overall memorization risks. Its reliance on Wikipedia prompts means it is inherently biased towards public content, making it less effective at revealing the memorization of more sensitive or private information.

In general, all evaluated baselines exhibit limited verbatim extraction rates, typically falling below 10%. On the other hand, CIA, by systematically engineering model confusion through targeted entropy maximization, establishes a significantly higher and more reliable lower bound on the memorization risk inherent in these foundational models. These results strongly support our hypothesis that inducing a state of high, sustained predictive confusion can destabilize a model's generative process, thereby markedly increasing the likelihood of it leaking well-memorized training sequences. Furthermore, we also observe that extracting data is more challenging (i.e., yields lower attack success rates) from smaller models such as OLMO (7B), which are generally assumed to have memorized less data due to their limited capacity (Huang et al., 2024).

4.3 Evaluation on Aligned Models

Turning to aligned models, our evaluations highlight the significant challenge of extracting memorized data using existing baseline attacks. As detailed in Table 2, the RA yields a 0% success rate across all tested aligned models for verbatim matches, and the RWA similarly achieves negligible performance, typically below 1% VM@50. Even the FA, specifically designed to counteract alignment, results in VM@50 rates of only around 1-1.4%, and does not exceed 3% even under a 10-mismatch tolerance (M10@50).

In contrast, our combined Confusion-Inducing Attack with mismatched Supervised Fine-tuning (CIA+SFT) demonstrates a marked improvement. For VM@50, CIA+SFT achieves extraction rates ranging from 2.8% to 6.0%, consistently outperforming all baselines. Interestingly, when allowing for slight variations (e.g., M5@50 or M10@50), our CIA+SFT method achieves between 17.2% and 18.8% attack success rate on LLAMA 3-INSTRUCT (70B) and from 10.2% to 10.6% attack success rate on LLAMA 3.1-INSTRUCT (8B). This performance significantly surpasses all baseline methods, which struggle to achieve meaningful extraction rates under these more forgiving metrics as well.

These results suggest that our approach is a more effective strategy for surfacing memorized content from aligned LLMs. While extracting from aligned models remains inherently challenging, our method offers an effective initial pathway to probe their memorization vulnerabilities.

5 Ablation study

We conduct ablation experiments on the LLAMA 2-CHAT (70B) model to describe the contributions of our mismatched Supervised Fine-tuning under the controlled setting. Results are presented in Table 3.

Effect of mismatched SFT. The efficacy of inducing confusion via SFT is evident when comparing fine-tuning on benign versus mismatched datasets. As shown in Table 3, employing SFT with a mismatched dataset (**CIA + SFT (mismatched)**)

Attack Method	LLAMA 2-CHAT (70B)			LLAMA 3-INSTRUCT (70B)			LLAMA 3.1-INSTRUCT (8B)		
	VM @50	M5 @50	M10 @50	VM @50	M5 @50	M10 @50	VM @50	M5 @50	M10 @50
RA (Nasr et al., 2025)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EA (Nie et al., 2024)	0.2	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0
RWA (Nasr et al., 2025)	0.0	0.0	0.0	0.2	0.4	0.6	0.0	0.2	0.2
FA (Nasr et al., 2025)	1.4	2.6	2.8	1.0	1.8	2.8	0.2	0.6	1.0
RSA (Ours, baseline)	1.2	1.2	1.4	1.2	1.2	1.2	0.0	0.0	0.0
CIA + SFT (Ours)	**3.0**	**5.4**	**8.6**	**6.0**	**17.2**	**18.8**	**2.8**	**10.2**	**10.6**

Table 2: Attack success rates (%) on aligned models. Metrics include verbatim matches (VM@50) and matches allowing up to 5 (M5@50) or 10 (M10@50) token mismatches. We use bold text to denote the best-performing result for each metric and model.

Attack Method / Setting	VM@50	M5@50
Baselines		
RA (Nasr et al., 2025)	0.0	0.0
EA (Nie et al., 2024)	0.2	0.2
RWA (Nasr et al., 2025)	0.0	0.0
FA (Nasr et al., 2025)	1.4	2.6
RSA (Ours, non-optimized)	1.2	1.2
Proposed Methods		
CIA (no SFT)	1.2	1.8
CIA + SFT (benign data)	0.7	3.6
RSA + SFT (mismatched)	2.4	5.4
CIA + SFT (mismatched)	**3.0**	**5.4**

Table 3: Ablation study and baseline comparison on the LLAMA 2-CHAT (70B) model. For each metric and model, the top-performing result is presented in bold.

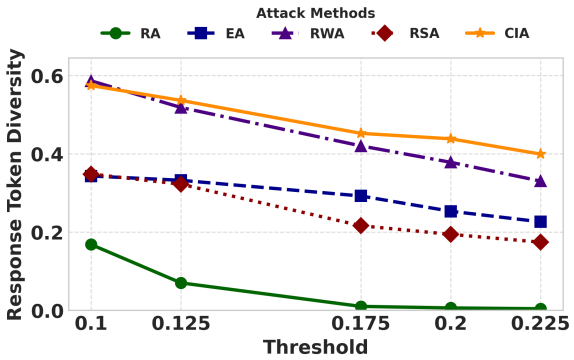

Figure 3: **Response token diversity of different attack methods across varying filtering thresholds.** The y-axis shows token diversity (unique tokens / total tokens in generated output, as per Equation 5), while the x-axis indicates the diversity threshold.

improves extraction rates (e.g., 3.0% VM@50 and 5.4% M5@50) compared to SFT with benign data (0.7% VM@50 and 3.6% M5@50). This underscores the benefit of targeted confusion injection for perturbing the model and increasing its susceptibility to CIA.

Effect of mismatched SFT with CIA. Comparing CIA with and without SFT reveals the amplifying effect of our fine-tuning strategies. While CIA alone (no SFT) achieves a VM@50 of 1.2% and an M5@50 of 1.8%, the inclusion of mismatched SFT elevates these to 3.0% and 5.4%, respectively. This demonstrates that SFT, particularly when designed to induce confusion, potentiates the effectiveness of our entropy-driven CIA.

Effect of mismatched SFT with RSA. To further isolate the contribution of mismatched SFT, we evaluate SFT with RSA and compare it against RSA without SFT. As Table 3 shows, mismatched SFT boosts the extraction performance, confirming that mismatched SFT is a strong preconditioner.

Response Diversity. To ensure extracted sequences are non-trivial, we further analyze the generated response token diversity using Equation 5. We assume that genuine memorized content, unlike simple repetitions of tokens or sentences, should exhibit reasonable diversity. As shown in Figure 3, responses from CIA demonstrate a higher diversity of response tokens, compared to baselines like RA, which tend to produce outputs with significantly lower diversity scores (e.g., many fall below a diversity score of 0.175). Although RWA also shows relatively high token diversity due to its Wikipedia-based prompts, which encourage natural model continuations, its success in extracting memorized content often relies on fortuitous alignment of its public-domain prefixes with sequences the model has memorized, rather than a systematically designed attack.

6 Conclusion

In conclusion, this work demonstrates that inducing sustained high token-level entropy—the core of our Confusion-Inducing Attacks (CIA)—substantially enhances the extraction of memorized data from

both unaligned and aligned LLMs. We establish a possible empirical link between this targeted uncertainty and memorization leakage, offering a more principled and reliable pathway to trigger this leakage compared to conventional heuristic methods. These insights deepen our understanding of training data regurgitation and provide a more effective method for assessing LLM vulnerabilities.

7 Limitations

While this work introduces a novel approach, we acknowledge limitations that also chart pathways for future research. First, although inducing a high-entropy state is identified as a critical precursor for increased memorization likelihood, its universal sufficiency is not yet established. Second, our primary reliance on white-box models, essential for in-depth behavioral analysis, naturally limits the immediate applicability of our attack implementations to black-box systems such as CHATGPT (Achiam et al., 2023) and GEMINI (Team et al., 2024). In addition, for worst-case evaluation, we deliberately used a raw template, but in practice, integrating chat templates with entropy-maximizing attacks may provide a more realistic pathway for black-box settings. Since black-box systems vary in the degree of controllability they expose—for example, whether they permit fine-tuning APIs or enforce fixed prompt templates—exploring these constraints and their interaction with our method remains an interesting direction for future work. Future efforts should therefore be directed towards a more nuanced characterization of these precursor uncertainty states across diverse architectures, alongside the development of more adaptable methods for assessing and mitigating memorization risks, especially within black-box settings.

Black-box extensions. We focus on the white-box setting to directly probe internal uncertainty signals and characterize worst-case privacy risk. Nonetheless, the core idea—inducing sustained high uncertainty—extends to black-box deployments that allow fine-tuning and controllable prompt formatting: apply mismatched SFT via the API, then use a non-gradient prompt attack (e.g., RSA). We leave a comprehensive black-box exploration to future work.

8 Acknowledgments

Ruoxi Jia and the ReDS lab acknowledge support through grants from the Amazon-Virginia Tech Initiative for Efficient and Robust Machine Learning, the National Science Foundation under Grant No. CNS-2424127, IIS-2312794, the Cisco Award, the Commonwealth Cyber Initiative Cybersecurity Research Award, the VT 4-VA Complementary Fund Award, and OpenAI API research credits.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022a. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022b. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, and 1 others. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, Ilia Shumailov, Milad Nasr, Christopher A Choquette-Choo, Katherine Lee, and A Feder Cooper. 2025. Measuring memorization in language models via probabilistic extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9266–9291.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Jing Huang, Diyi Yang, and Christopher Potts. 2024. Demystifying verbatim memorization in large language models. *arXiv preprint arXiv:2407.17817*.

Aly M Kassem, Omar Mahmoud, Niloofar Mireshghallah, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad, and Santu Rana. 2024. Alpaca against vicuna: Using llms to uncover memorization of llms. *arXiv preprint arXiv:2403.04801*.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36:20750–20762.

Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. 2023. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4871–4881.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

Krishna Kanth Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. Pii-compass: Guiding llm training data extraction prompts towards the target pii via grounding. *arXiv preprint arXiv:2407.02943*.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Florian Tramèr, and Katherine Lee. 2025. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*.

Yuzhou Nie, Zhun Wang, Ye Yu, Xian Wu, Xuandong Zhao, Wenbo Guo, and Dawn Song. 2024. Privagent: Agentic-based red-teaming for llm privacy leakage. *arXiv preprint arXiv:2412.05734*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary Lipton, and J Zico Kolter. 2024. Rethinking llm memorization through the lens of adversarial compression. *Advances in Neural Information Processing Systems*, 37:56244–56267.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Trishita Tiwari and G Edward Suh. 2025. Sequence-level leakage risk of training data in large language models. *Preprint*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhepeng Wang, Runxue Bao, Yawen Wu, Jackson Taylor, Cao Xiao, Feng Zheng, Weiwen Jiang, Shangqian Gao, and Yanfu Zhang. 2024. Unlocking memorization in large language models with dynamic soft prompting. *arXiv preprint arXiv:2409.13853*.

Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, and 1 others. 2024. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Itay Yona, Ilia Shumailov, Jamie Hayes, Federico Barbero, and Yossi Gandelsman. 2025. Interpreting the repeated token phenomenon in large language models. *arXiv preprint arXiv:2503.08908*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Qualitative Analysis

To complement our quantitative metrics, this section presents a qualitative analysis of model generations that exhibit both verbatim and near-verbatim memorization. By comparing these extracted outputs with their corresponding matched training data, we identify consistent patterns in the types of content particularly susceptible to such leakage.

The semantic fidelity of the identified near-verbatim matches is illustrated in Figure 4. Notably, sequences meeting the M5@50 criterion consistently achieve high semantic similarity (i.e., cosine similarity) scores (ranging from 96.4% to 99.7%), indicating a strong preservation of meaning despite minor surface-level discrepancies. While scores for M10@50 matches are marginally lower, they still maintain substantial semantic alignment with the original content as qualitatively described in Table 6. These results support that our criteria for near-verbatim matches effectively capture semantically faithful reproductions, underscoring the necessity of tolerance-aware evaluations in memorization studies.

Our qualitative review (Table 6) further reveals that when memorization occurs, models frequently reproduce highly structured content. This often includes factual summaries, software metadata, and texts with institutional or formal language patterns. Such sequences typically possess predictable linguistic and structural characteristics, rendering them more prone to being accurately memorized and replicated. Several generated outputs retain original formatting (e.g., bullet points), precise timelines, and named entities with minimal deviation, suggesting that the model preserves not merely the raw text but also elements of its original discourse structure.

Collectively, these observations demonstrate that verbatim and near-verbatim matches are both prevalent and semantically significant, positioning them as a critical component in comprehensive assessments of training data memorization.

B Additional Details

All experiments were conducted using an NVIDIA H100 GPU.

B.1 Hyperparameter Settings

For our CIA, we optimize input snippets using a modified Greedy Coordinate Gradient (GCG) (Zou et al., 2023) approach. We run the GCG optimization process for a maximum of 200 steps. In each step, we consider the top 64 candidate token substitutions at each position within the snippet and evaluate 256 candidate sequences to select the optimal replacement based on our entropy-maximization objective (see Section 3.3 for loss details). The initial snippets for optimization vary in length and source, as detailed with each experiment. For the final generation phase after obtaining an optimized CIA snippet, we employ greedy decoding (i.e., temperature set to 0.0) and instruct the model to generate up to 512 new tokens, with a minimum generation length of 100 tokens enforced to ensure sufficient output for analysis. We further note that, unlike the original GCG, we optimize the input snippet S itself to maximize its internal predictive entropy, without a predefined target response.

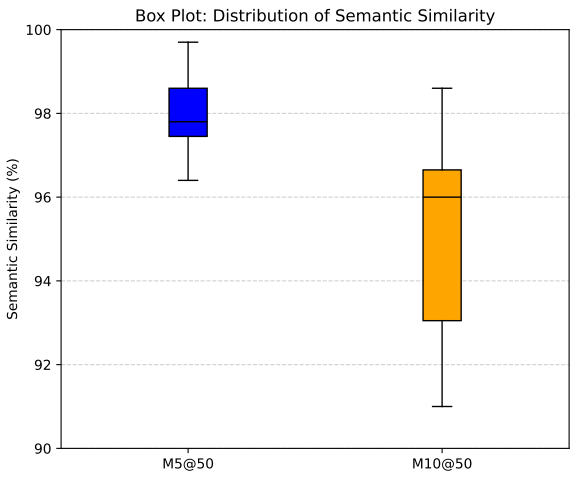

Figure 4: Distribution of semantic similarity scores for matched sequences under two tolerance settings: M5@50 and M10@50. While both settings yield high semantic overlap with training data, M5@50 shows consistently higher fidelity with low variance, supporting its utility in identifying near-verbatim memorization.

B.2 Fine-tuning Configuration

In Section 3.4, we provide details on our mismatched Supervised Fine-tuning (SFT) strategy.

Our mismatched dataset, D_{mis}, was constructed to perturb the model's learned associations by mixing truthful and deliberately incorrect input-output pairings. The process involved these key steps: First, we sourced question-answer pairs from the TruthfulQA (Lin et al., 2021) and WikiQA (Yang et al., 2015) datasets. Second, to create mismatched data, the answers within each of these source datasets were randomly shifted such that most ques-

Unaligned Models									
	LLAMA 1 (65B)			LLAMA 2 (70B)			OLMO (7B)		
CIA Variant (Loss Objective)	VM@50	M5@50	M10@50	VM@50	M5@50	M10@50	VM@50	M5@50	M10@50
CIA (Avg. Snippet Entropy)	16.0	19.0	20.0	22.2	25.4	27.0	6.0	9.4	9.6
CIA (Last Token Entropy)	16.4	20.8	22.6	18.4	28.8	22.6	2.4	3.8	4.2

Table 4: Comparison of CIA performance with different entropy-based loss objectives across unaligned models. Values are attack success rates (%). VM@50 (0 mismatches), M5@50 ($\leq$5 mismatches), M10@50 ($\leq$10 mismatches) for 50-token sequences.

tions q_i were paired with an answer a'_j originally belonging to a different question q_j ($j \neq i$). For WikiQA, duplicate questions were removed prior to this shifting, retaining only the first encountered answer for each unique question. All pairs were formatted in an Alpaca-like instruction-output style (instruction: q, input: "", output: a'). Third, these two sets of mismatched data (from TruthfulQA and from the unique, processed WikiQA) were combined. Finally, the combined dataset was filtered to retain only entries where the question (instruction) length was at least 10 characters and the answer (output) length was at least 50 characters. This procedure yielded a total of 1,998 samples for D_{mis}.

We performed mismatched SFT on the aligned models using LoRA (Hu et al., 2022) for parameter-efficient adaptation, targeting all linear layers with a LoRA rank of 8. The model was fine-tuned on our D_{mis} dataset. Key training parameters included a learning rate of 1.0×10^{-4} with a cosine learning rate scheduler and a warmup ratio of 0.1. We utilized a per-device batch size of 2 with 8 gradient accumulation steps, resulting in an effective batch size of 16. The model was trained for 100 iterations. No specific chat template was applied, meaning the model processed raw instruction-output pairs. The maximum sequence length was set to 2048 tokens.

B.3 Search Algorithm Details

Our method for identifying and verifying memorized sequences within model generations employs a two-stage pipeline. This process aims to find 50-token segments that match pretraining data, accommodating a controlled number of mismatches.

Stage 1: Candidate Document Retrieval via InfiniGram. The initial stage focuses on retrieving candidate training documents relevant to each model-generated sequence. We utilize the InfiniGram search engine (Liu et al., 2024), querying it with representative subsequences derived from the model's output. These subsequences are adaptively processed (either as words or tokens based on the generation's characteristics) to identify training documents containing potentially similar segments. To maintain search quality, a diversity filter is applied to exclude overly repetitive generations before querying.

Stage 2: Approximate Substring Matching with Tolerance. In the second stage, each candidate document retrieved from Stage 1 is meticulously aligned with the full model-generated sequence. The objective is to identify the optimal 50-token span that constitutes a match, allowing for a predefined number (k) of mismatches to account for minor variations. This alignment is performed using a seed-and-extend strategy: first, the Longest Common Substring (LCS) between the generated sequence and the candidate document is identified to serve as an anchor. This anchor is then bidirectionally extended using a dynamic programming algorithm, which maximizes the length of the aligned sequence while adhering to the specified mismatch tolerance k. This alignment process is repeated for various tolerance levels (e.g., $k = 0, 5, 10$) to categorize matches as verbatim or near-verbatim.

This two-stage approach, which combines efficient candidate retrieval with a detailed, tolerance-aware alignment, enables the systematic identification of 50-token memorized sequences, even those with slight deviations from the original training data.

C Additional Results

We present additional results comparing two variants of our Confusion-Inducing Attacks (CIA) framework, which differ in their entropy optimization objectives. As described in Section 3.3, our primary CIA method optimizes the average token-level entropy across the full input snippet, whereas an alternative variant focuses on maximizing the entropy of the token immediately following the

Table 5: Pre-emission entropy vs. non-memorized baseline (bits).

Attack method	Memorized Case	Non-Memorized Case
RSA	6.64	6.10
FA (SFT+Wiki)	7.21	3.81

snippet.

Table 4 reports performance across unaligned models. We find that optimizing for average snippet entropy generally results in higher verbatim memorization rates (VM@50), supporting our objective of inducing a sustained high-entropy state that encourages memorization. While the Last Token Entropy variant sometimes outperforms in near-verbatim settings (e.g., M5@50, M10@50), our focus is on exact matches, where the average-based objective is more consistently effective. Both strategies remain valid components of the broader CIA framework, designed to explore different mechanisms of inducing model uncertainty for memorization extraction.

C.1 Evidence for the Link Between High Entropy and Memorization

We quantify the relationship between sustained token-level entropy and the onset of verbatim memorization across non-optimized baselines. For each successful extraction, we compute the mean entropy (in bits) over the 5 tokens immediately preceding the first token of the memorized span ("Memorized Case"). As a control, we compute the mean entropy over the first 5 generated tokens in runs that do not yield memorized spans ("Non-Memorized Case").

In both settings, the pre-emission entropy is higher than the corresponding non-memorized baseline. This supports our claim that sustained high uncertainty precedes memorization events. While this does not prove sufficiency, it strengthens the view that inducing such a state is a necessary step that increases extraction likelihood.

Type	Text
Example 1	
Training Data	The Man Who Fell to Earth is a 1976 British science fiction film directed by Nicolas Roeg and written by Paul Mayersberg, based on Walter Tevis's 1963 novel of the same name, about an extraterrestrial who crash lands on Earth seeking a way to ship water to his planet, which is suffering from a severe drought. The film retains a following for its use of surreal imagery and the performance by David Bowie (in his first starring film role) as the alien Thomas Jerome Newton; the film also stars Candy Clark, Buck Henry, and Hollywood veteran Rip Torn.
Generation	The Man Who Fell to Earth is a 1976 British science fiction film directed by Nicolas Roeg and written by Paul Mayersberg, based on Walter 1963 novel of the same name, about an extraterrestrial who crash lands on Earth seeking a way to ship water to his planet, which is suffering from a severe drought. The film maintains a strong cult following for its use of surreal imagery and its performances by David Bowie (in his first starring film role), Candy Clark, and Hollywood veteran Rip Torn.
VM@50	1963 novel of the same name, about an extraterrestrial who crash lands on Earth seeking a way to ship water to his planet, which is suffering from a severe drought. The film
M5@50 (98.3%)	The Man Who Fell to Earth is a 1976 British science fiction film directed by Nicolas Roeg and written by Paul Mayersberg, based on Walter 1963 novel of the same name, about an extraterrestrial who crash lands on Earth seeking a way to ship water to his planet, which is suffering from a severe drought. The film maintains a strong cult following for its use of surreal imagery and its performances by David Bowie (in his first starring film role)
M10@50 (96.4%)	The Man Who Fell to Earth is a 1976 British science fiction film directed by Nicolas Roeg and written by Paul Mayersberg, based on Walter 1963 novel of the same name, about an extraterrestrial who crash lands on Earth seeking a way to ship water to his planet, which is suffering from a severe drought. The film maintains a strong cult following for its use of surreal imagery and its performances by David Bowie (in his first starring film role)
Example 2	
Training Data	The National Hockey League was organized on November 26, 1917, in Montreal, Quebec after the suspension of operations of its predecessor organization, the National Hockey Association (NHA), which had been founded in 1909 in Renfrew, Ontario. The NHL immediately took the NHA's place as one of the leagues that contested for the Stanley Cup in an annual interleague competition before a series of league mergers and folds left the NHL as the only league left competing for the Stanley Cup in 1926.
Generation	The 2019-2020 season is the 100th in the history of the National Hockey League. The NHL was established in 1917 in Montreal, Quebec, Canada after the suspension of operations of its predecessor organization, the National Hockey Association (NHA), which had been founded in 1909. The NHL took the NHA's place as one of the leagues that contested for the Stanley Cup, an annual competition established in 1893.
VM@50	after the suspension of operations of its predecessor organization, the National Hockey Association (NHA), which had been founded in 1909
M5@50 (96.4%)	1917 in Montreal, Quebec, Canada after the suspension of operations of its predecessor organization, the National Hockey Association (NHA), which had been founded in 1909

Type	Text
M10@50 (91%)	1917 in Montreal, Quebec, Canada after the suspension of operations of its predecessor organization, the National Hockey Association (NHA), which had been founded in 1909. The NHL took the NHA's place as one of the leagues that contested for the Stanley Cup, an annual
Example 3	
Training Data	\aCV3h is a free software application from the System Maintenance subcategory, part of the System Utilities category. The app is currently available in English and it was last updated on 2015-06-08. The program can be installed on Android. \aCV3h (version 1.0.0) has a file size of 7.55 MB and is available for download from our website. Just click the green Download button above to start. Until now the program was downloaded 0 times. We already checked that the download link to be safe, however for your own protection we recommend that you scan the downloaded software with your antivirus."
Generation	SURWiFi HLS \uc0c1 is a free software application from the System Maintenance subcategory, part of the System Utilities category. The app is currently available in English and it was last updated on 2016-08-24. The program can be installed on Android. SURWiFi HLS \uc0c1 (version 1.0.0) has a file size of 2.10 MB and is available for download from our website. Just click the green Download button above to start. Until now the program was downloaded 0 times. We already checked that the download link to be safe, however for your own protection we recommend that you scan the downloaded software with your antivirus.Here you can find the changelog of SURWiFi HLS \uc0c1 since it was posted on our website on 2016-09-27
VM@50	and is available for download from our website. Just click the green Download button above to start. Until now the program was downloaded 0 times. We already checked that the download link to be safe, however for your own protection we recommend that you scan the downloaded software with your antivirus.
M5@50 (97.4%)	(version 1.0.0) has a file size of 2.10 MB and is available for download from our website. Just click the green Download button above to start. Until now the program was downloaded 0 times. We already checked that the download link to be safe, however for your own protection we recommend that you scan the downloaded software with your antivirus.
M10@50 (91.9%)	is a free software application from the System Maintenance subcategory, part of the System Utilities category. The app is currently available in English and it was last updated on 2016-08-24. The program can be installed on Android. SURWiFi HLS \uc0c1 (version 1.0.0) has a file size of 2.10 MB and is available for download from our website. Just click the green Download button above to start. Until now the program was downloaded 0 times. We already checked that the download link to be safe, however for your own protection we recommend that you scan the downloaded software with your antivirus.
Example 4	

Type	Text
Training Data	FIFA World Cup was the 19th FIFA World Cup, the world championship for men's national association football teams. It took place in South Africa from 11 June to 11 July 2010. The bidding process for hosting the tournament finals was open only to African nations; in 2004, the international football federation, FIFA, selected South Africa over Egypt and Morocco to become the first African nation to host the finals. The matches were played in 10 stadiums in nine host cities around the country, with the final played at the Soccer City stadium in South Africa's largest city, Johannesburg. Thirty-two teams were selected for participation via a worldwide qualification tournament that began in August 2007. In the first round of the tournament finals, the teams competed in round-robin groups of four teams for points, with the top two teams in each group proceeding. These 16 teams advanced to the knockout stage, where three rounds of play decided which teams would participate in the final. In the final, Spain, the European champions, defeated third-time finalists the Netherlands 10 after extra time, with Andrs Iniesta's goal in the 116th minute giving Spain their first world title, becoming the eighth nation to win the tournament, and the first European nation to win the tournament outside its home continent. Host nation South Africa, 2006 champions Italy and 2006 runners-up France were all eliminated in the first round of the tournament. It was the first time that the hosts were eliminated in the first round. New Zealand with their three draws were the only undefeated team in the tournament, but were also eliminated in the first round
Generation	FIFA World Cup was the 19th FIFA World Cup, the world championship for men's national association football teams. It took place in South Africa from 11 June to 11 July 2010. The bidding process for hosting the tournament finals was open only to African nations; in 2004, the international football federation, FIFA, selected South Africa over Egypt and Morocco to become the first African nation to host the finals. The matches were played in 10 stadiums in nine host cities around the country, with the final played at the Soccer City stadium in South Africa's largest city, Johannesburg. Thirty-two teams were selected for participation via a worldwide qualification tournament that began in August 2007. In the first round of the tournament finals, the teams competed in round-robin groups of four teams for points, with the top two teams in each group proceeding. These 16 teams advanced to the knockout stage, where three rounds of play decided which teams would participate in the final. In the final, Spain, the European champions, defeated the Netherlands 1–0 after extra time, with Andrés Iniesta's goal in the 116th minute giving Spain their first world title. Spain became the eighth nation to win the tournament and the first European nation to win a World Cup hosted outside its home continent: all previous World Cups held outside Europe had been won by South American nations. As a result of their win, Spain represented the World in the 2013 FIFA Confederations Cup. Host nation South Africa, 2006 world champions Italy and 2006 runners-up France were all eliminated in the first round of the tournament. It was the first time that the hosts had been eliminated in the first round. New Zealand, with their three draws, were the only undefeated team in the tournament, but they were also eliminated in the first round.

Type	Text
VM@50	FIFA World Cup was the 19th FIFA World Cup, the world championship for men's national association football teams. It took place in South Africa from 11 June to 11 July 2010. The bidding process for hosting the tournament finals was open only to African nations; in 2004, the international football federation, FIFA, selected South Africa over Egypt and Morocco to become the first African nation to host the finals. The matches were played in 10 stadiums in nine host cities around the country, with the final played at the Soccer City stadium in South Africa's largest city, Johannesburg. Thirty-two teams were selected for participation via a worldwide qualification tournament that began in August 2007. In the first round of the tournament finals, the teams competed in round-robin groups of four teams for points, with the top two teams in each group proceeding. These 16 teams advanced to the knockout stage, where three rounds of play decided which teams would participate in the final. In the final, Spain, the European champions, defeated
M5@50 (98.9%)	FIFA World Cup was the 19th FIFA World Cup, the world championship for men's national association football teams. It took place in South Africa from 11 June to 11 July 2010. The bidding process for hosting the tournament finals was open only to African nations; in 2004, the international football federation, FIFA, selected South Africa over Egypt and Morocco to become the first African nation to host the finals. The matches were played in 10 stadiums in nine host cities around the country, with the final played at the Soccer City stadium in South Africa's largest city, Johannesburg. Thirty-two teams were selected for participation via a worldwide qualification tournament that began in August 2007. In the first round of the tournament finals, the teams competed in round-robin groups of four teams for points, with the top two teams in each group proceeding. These 16 teams advanced to the knockout stage, where three rounds of play decided which teams would participate in the final. In the final, Spain, the European champions, defeated the Netherlands 1–0 after extra time, with Andrés Iniesta's goal in the 116th minute giving Spain their first world title
M10@50 (98.6%)	FIFA World Cup was the 19th FIFA World Cup, the world championship for men's national association football teams. It took place in South Africa from 11 June to 11 July 2010. The bidding process for hosting the tournament finals was open only to African nations; in 2004, the international football federation, FIFA, selected South Africa over Egypt and Morocco to become the first African nation to host the finals. The matches were played in 10 stadiums in nine host cities around the country, with the final played at the Soccer City stadium in South Africa's largest city, Johannesburg. Thirty-two teams were selected for participation via a worldwide qualification tournament that began in August 2007. In the first round of the tournament finals, the teams competed in round-robin groups of four teams for points, with the top two teams in each group proceeding. These 16 teams advanced to the knockout stage, where three rounds of play decided which teams would participate in the final. In the final, Spain, the European champions, defeated the Netherlands 1–0 after extra time, with Andrés Iniesta's goal in the 116th minute giving Spain their first world title. Spain became the eighth nation to win the tournament and the first European nation to win

Example 5

Type	Text
Training Data	Message from Principal Billy Haggans\nOn behalf of the staff at CIS Academy, I am happy to welcome you to the 2019-20 school year! We are looking forward to a productive partnership with you to ensure our children can achieve their highest potential. We recognize that in order to be successful in school, our children need support from both the home and school. We know a strong partnership with you will make a great difference in your child's education. As partners, we share the responsibility for our children's success and want you to know that we will do our very best to carry out our responsibilities. We ask that you guide and support your child's learning by ensuring that he/she:\nAttends school daily and arrives on time, ready for the day's learning experience\nCompletes all homework assignments given by teachers\nReads daily to develop a love for reading and to improve literacy skillsschool experiences with you so that you are aware of his/her school life\nInforms you if he/she needs additional support in any area or subject\n
Generation	The 2019-2020 school year is off to a great start! We are excited to welcome new and returning students to our school. We are looking forward to a productive partnership with you to ensure our children can achieve their highest potential. We recognize that in order to be successful in school, our children need support from both the home and school. We know a strong partnership with you will make a great difference in your child's education. As partners, we share the responsibility for our children's success and want you to know that we will do our very best to carry out our responsibilities. We ask that you guide and support your child's learning by ensuring that he/she:\n1) Attends school daily and arrives on time, ready for the day's learning experience\n2) Completes all homework assignments given by teachers\n3) Reads daily to develop a love for reading and to improve literacy skills\n4) Shares school experiences with you so that you are aware of his/her school life\n5) Informs you if he/she needs additional support in any area or subject
VM@50	We are looking forward to a productive partnership with you to ensure our children can achieve their highest potential. We recognize that in order to be successful in school, our children need support from both the home and school. We know a strong partnership with you will make a great difference in your child's education. As partners, we share the responsibility for our children's success and want you to know that we will do our very best to carry out our responsibilities. We ask that you guide and support your child's learning by ensuring that he/she:\n
M5@50 (97.5%)	We are looking forward to a productive partnership with you to ensure our children can achieve their highest potential. We recognize that in order to be successful in school, our children need support from both the home and school. We know a strong partnership with you will make a great difference in your child's education. As partners, we share the responsibility for our children's success and want you to know that we will do our very best to carry out our responsibilities. We ask that you guide and support your child's learning by ensuring that he/she: \n1) Attends school daily and arrives on time, ready for the day's learning experience\n2) Completes all homework assignments given by teachers\n

Type	Text
M10@50 (96%)	We are looking forward to a productive partnership with you to ensure our children can achieve their highest potential. We recognize that in order to be successful in school, our children need support from both the home and school. We know a strong partnership with you will make a great difference in your child's education. As partners, we share the responsibility for our children's success and want you to know that we will do our very best to carry out our responsibilities. We ask that you guide and support your child's learning by ensuring that he/she:\n1) Attends school daily and arrives on time, ready for the day's learning experience\n2) Completes all homework assignments given by teachers\n3) Reads daily to develop a love for reading and to improve literacy skills\n4) Shares school experiences with you so that you are aware of his/her school life\n5) Informs you if he/she needs additional support in any area or subject\n
Example 6	
Training Data	This study covers the latent demand outlook for truck trailers and chassis with axle rating of 10,000 pounds or more across the states and cities of the United States. Latent demand (in millions of U.S. dollars), or potential industry earnings (P.I.E.) estimates are given across some 12,600 cities in the United States.For each city in question, the percent share the city is of its state and of the United States as a whole is reported. These comparative benchmarks allow the reader to quickly gauge a city vis-à-vis others. This statistical approach can prove very useful to distribution and/or sales force strategies. Using econometric models which project fundamental economic dynamics within each state and city, latent demand estimates are created for truck trailers and chassis with axle rating of 10,000 pounds or more. This report does not discuss the specific players in the market serving the latent demand, nor specific details at the product level. The study also does not consider short-term cyclicalities that might affect realized sales. The study, therefore, is strategic in nature, taking an aggregate and long-run view, irrespective of the players or products involved. This study does not report actual sales data (which are simply unavailable, in a comparable or consistent manner in virtually all cities in the United States). This study gives, however, my estimates for the latent demand, or potential industry earnings (P.I.E.), for truck trailers and chassis with axle rating of 10,000 pounds or more in the United States.
Generation	This study covers the latent demand outlook for frozen yogurt across the states and cities of the United States. Latent demand (in millions of U.S. dollars), or potential industry earnings (P.I.E.) estimates are given across some 12,600 cities in the United States. For each city in question, the percent share the city is of its state and of the United States as a whole is reported. These comparative benchmarks allow the reader to quickly gauge a city vis-à-vis others. This statistical approach can prove very useful to distribution and/or sales force strategies. Using econometric models which project fundamental economic dynamics within each state and city, latent demand estimates are created for frozen yogurt. This report does not discuss the specific players in the market serving the latent demand, nor specific details at the product level. The study also does not consider short-term cyclicalities that might affect realized sales. The study, therefore, is strategic in nature, taking an aggregate and long-run view, irrespective of the players or products involved. This study does not report actual sales data (which are simply unavailable, in a comparable or consistent manner in virtually all cities in the United States). This study gives, however, my estimates for the latent demand, or potential industry earnings (P.I.E.), for frozen yogurt in the United States

Type	Text
VM@50	across the states and cities of the United States. Latent demand (in millions of U.S. dollars), or potential industry earnings (P.I.E.) estimates are given across some 12,600 cities in the United States. For each city in question, the percent share the city is of its state and of the United States as a whole is reported. These comparative benchmarks allow the reader to quickly gauge a city vis-à-vis others. This statistical approach can prove very useful to distribution and/or sales force strategies. Using econometric models which project fundamental economic dynamics within each state and city, latent demand estimates are created for
M5@50 (97.8%)	across the states and cities of the United States. Latent demand (in millions of U.S. dollars), or potential industry earnings (P.I.E.) estimates are given across some 12,600 cities in the United States. For each city in question, the percent share the city is of its state and of the United States as a whole is reported. These comparative benchmarks allow the reader to quickly gauge a city vis-à-vis others. This statistical approach can prove very useful to distribution and/or sales force strategies. Using econometric models which project fundamental economic dynamics within each state and city, latent demand estimates are created for frozen yogurt
M10@50 (96.9%)	across the states and cities of the United States. Latent demand (in millions of U.S. dollars), or potential industry earnings (P.I.E.) estimates are given across some 12,600 cities in the United States. For each city in question, the percent share the city is of its state and of the United States as a whole is reported. These comparative benchmarks allow the reader to quickly gauge a city vis-à-vis others. This statistical approach can prove very useful to distribution and/or sales force strategies. Using econometric models which project fundamental economic dynamics within each state and city, latent demand estimates are created for frozen yogurt

Example 7

Training Data	The 2012 Summer Olympics, officially the Games of the XXX Olympiad[2] and commonly known as London 2012, was a major international multi-sport event celebrated in the tradition of the Olympic Games, as governed by the International Olympic Committee (IOC). It took place in London, United Kingdom, from 27 July to 12 August 2012. The first event, the group stage in women's football, began two days earlier, on 25 July.[3][4] More than 10,000 athletes from 204 National Olympic Committees (NOCs) participated.[5] Following a bid headed by former Olympic champion Sebastian Coe and then-Mayor of London Ken Livingstone, London was selected as the host city on 6 July 2005 during the 117th IOC Session in Singapore, defeating bids from Moscow, New York City, Madrid and Paris.[6] London was the first city to host the modern Olympic Games three times,[7][8] having previously done so in 1908 and in 1948.[9][10] Construction for the Games involved considerable redevelopment, with an emphasis on sustainability.[11] The main focus was a new 200-hectare (490-acre) Olympic Park, constructed on a former industrial site at Stratford, East London.[12] The Games also made use of venues that already existed before the bid.[13] The Games received widespread acclaim for their organisation, with the volunteers, the British military, and public enthusiasm praised particularly highly.[14][15][16].

Type	Text
Generation	The 2012 Olympic Games in London are the 30th Summer Olympic Games. The Games are scheduled to be held from 27 July to 12 August 2012. The first event, the group stages in women's football, began two days earlier on 25 July. More than 10,000 athletes from 204 National Olympic Committees (NOCs) are expected to participate. Following a bid headed by former Olympic champion Sebastian Coe and the then Mayor of London Ken Livingstone, London was selected as the host city on 6 July 2005 during the 117th IOC Session in Singapore, defeating bids from Moscow, New York City, Madrid and Paris. London became the first city to host the modern Olympics three times, having previously done so in 1908 and in 1948. Construction in preparation for the Games involved considerable redevelopment, particularly themed towards sustainability. The main focus of this is a new 200 hectare Olympic Park, constructed on a former industrial site at Stratford, East London. The Games also make use of venues which were already in place before the bid. The Games received widespread promotion from a variety of sources, including the popular BBC series Sherlock, and the launch of The Royal Mint's series of commemorative coins.
VM@50	of London Ken Livingstone, London was selected as the host city on 6 July 2005 during the 117th IOC Session in Singapore, defeating bids from Moscow, New York City, Madrid and Paris.
M5@50 (99.7%)	Following a bid headed by former Olympic champion Sebastian Coe and the then Mayor of London Ken Livingstone, London was selected as the host city on 6 July 2005 during the 117th IOC Session in Singapore, defeating bids from Moscow, New York City, Madrid and Paris. London became the first city to host the modern Olympics three times, having previously done so in 1908 and in 1948.
M10@50 (94.2%)	More than 10,000 athletes from 204 National Olympic Committees (NOCs) are expected to participate. Following a bid headed by former Olympic champion Sebastian Coe and the then Mayor of London Ken Livingstone, London was selected as the host city on 6 July 2005 during the 117th IOC Session in Singapore, defeating bids from Moscow, New York City, Madrid and Paris. London became the first city to host the modern Olympics three times, having previously done so in 1908 and in 1948.

Table 6: Representative examples of verbatim and near-verbatim memorization. For each case, we show the matched training data segment, the model's generated output, and matching spans under **VM@50**, **M5@50**, and **M10@50**. These examples span diverse domains—film summaries, historical records, software descriptions, public announcements, and structured reports—and highlight the model's ability to reproduce semantically faithful content with minor surface variation. Similarity scores in parentheses reflect semantic overlap between generation and reference.