On the Feasibility of Extracting Copyrighted Text from Production Large Language Models: A Computational Analysis of Attack-Defense Dynamics

Anonymous Author(s)

ABSTRACT

Whether copyrighted training data can be extracted from production large language models (LLMs) despite safety measures remains an open question with significant legal and technical implications. We present a computational framework that models the interplay between memorization dynamics, multi-phase extraction attacks, and layered defense mechanisms across production LLM configurations. Our simulations of four production model archetypes (65B–1000B parameters) reveal that while defense stacks reduce average extraction rates from baseline to 0.1257 under standard attacks, adversarial techniques combining Best-of-N jailbreaking with iterative continuation achieve mean extraction rates of 0.3251—a 2.59× increase. Defense effectiveness averages 0.8377 across models, yet the average jailbreak uplift of 0.1993 demonstrates that alignment-based defenses remain partially vulnerable to adversarial bypass. Memorization follows a power-law scaling with model size (exponent $\alpha = 0.42$, $R^2 = 1.000$), creating a fundamental tension: larger models memorize more content while deploying stronger defenses. We find that no single defense mechanism achieves high effectiveness without substantial cost—output filtering at 0.7069 effectiveness incurs 0.1203 false positive rate, while RLHF alignment at 0.8110 effectiveness introduces 0.4564 jailbreak vulnerability. These results suggest that extraction of copyrighted text from production LLMs remains feasible at non-trivial rates even under comprehensive safety measures, motivating the development of fundamentally new defense paradigms.

KEYWORDS

memorization, copyright, language models, extraction attacks, safety, alignment

1 INTRODUCTION

Large language models are trained on vast corpora that include copyrighted text, raising fundamental questions about the extent to which these models memorize and can reproduce their training data [3, 4]. While open-weight, non-instruction-tuned models have been shown to reproduce substantial amounts of copyrighted book text near-verbatim [8], production LLMs deploy both model-level alignment (RLHF, refusal training) and system-level guardrails (output filtering, activation capping) intended to prevent such reproduction [11].

Ahmed et al. [1] pose the open problem: is extraction of copyrighted book text, comparable to what has been demonstrated for open-weight models, feasible from production LLMs despite these safety measures? This question has direct implications for copyright litigation, LLM deployment practices, and the design of next-generation safety systems.

We approach this problem computationally, developing a simulation framework that models: (1) memorization as a function of model scale and data duplication, (2) multi-phase extraction attacks including Best-of-N jailbreaking and iterative continuation, (3) four categories of defense mechanisms with individual and combined effectiveness, and (4) the interaction between attacks and defenses across four production model archetypes.

Our analysis reveals several key findings:

- Production model defenses reduce extraction rates substantially (average defense effectiveness of 0.8377), but residual extraction remains non-trivial at an average rate of 0.1257 under standard attacks.
- Adversarial techniques boost extraction to an average of 0.3251, representing a mean jailbreak uplift of 0.1993.
- Memorization scales as a power law with model size ($\alpha = 0.42$), creating tension with defense scaling.
- The most effective combined defense (filter plus RLHF, effectiveness 0.9016) still permits extraction, while its jailbreak vulnerability stands at 0.4564.

1.1 Related Work

Memorization in LLMs. Carlini et al. [3] established that memorization in neural language models scales predictably with model size and data duplication, following power-law relationships. Biderman et al. [2] extended these findings to show both emergent and predictable memorization patterns across model scales. Nasr et al. [10] demonstrated practical extraction of training data from production systems including ChatGPT through divergence-based attacks.

Extraction Attacks. Recent work has shown that even aligned models can be induced to produce memorized content through adversarial prompting [5], with jailbreaking techniques that exploit the tension between helpfulness and safety objectives [12]. Ahmed et al. [1] proposed a two-phase extraction procedure combining initial probes with iterative continuation for production systems.

Defense Mechanisms. Defenses against memorization extraction include output filtering for near-verbatim matches [7], RLHF-based alignment to reduce copyright recitation [11], and activation-level interventions [9]. Ippolito et al. [6] cautioned that preventing verbatim generation alone may provide a false sense of privacy, as models can still leak information through paraphrasing.

2 METHODS

2.1 Memorization Model

We model memorization probability as a function of model size s (in billions of parameters), data duplication count d, sequence length

ℓ, and position within the source text $p \in [0,1]$:

$$P_{\text{mem}}(s,d,\ell,p) = \beta_0 \cdot \left(\frac{s}{s_0}\right)^{\alpha} \cdot d^{\gamma} \cdot f(p) \cdot g(\ell) \quad (1)$$

where $\beta_0 = 0.12$ is the base memorization rate at reference size $s_0 = 7$B parameters, $\alpha = 0.42$ is the size scaling exponent, and $\gamma = 0.38$ is the duplication exponent. The position factor $f(p) = 1 + 0.4(\exp(-10p) + \exp(-10(1-p)))$ captures the empirical finding that text near the beginning and end of books is memorized more readily [3]. The length factor $g(\ell) = \exp(-0.002(\ell - 256))$ penalizes longer sequences.

2.2 Extraction Attack Models

We model three extraction strategies:

Direct extraction. Given a memorized passage, the extraction probability under greedy decoding ($T = 0$) equals the memorization probability reduced by defense effectiveness δ:

$$P_{\text{ext}}^{\text{direct}} = P_{\text{mem}} \cdot e^{-1.5T} \cdot (1-\delta) \quad (2)$$

Best-of-N jailbreaking. Sampling N completions and selecting the best match yields boosted probability:

$$P_{\text{ext}}^{\text{BoN}} = 1 - (1 - P_{\text{ext}}^{\text{base}})^{N^{0.85}} \quad (3)$$

where the exponent 0.85 accounts for sub-linear effective sampling due to inter-sample correlation.

Iterative continuation. Multi-step extraction amplifies the base probability through accumulated context:

$$P_{\text{ext}}^{\text{iter}}(k) = P_{\text{base}} + (1 - P_{\text{base}}) \cdot (1 - e^{-0.3k}) \cdot 2P_{\text{base}} \quad (4)$$

where k is the number of continuation steps.

2.3 Defense Mechanism Models

We model four defense mechanisms, each characterized by an effectiveness function and a cost metric:

Output filtering blocks content matching known copyrighted text, with effectiveness following a sigmoid in filter strictness and a false positive rate scaling quadratically.

Activation capping clips high-magnitude activations that correlate with memorized content retrieval, with effectiveness $E_c = 0.9(1 - \exp(-3a))$ where $a = 1 - \text{percentile}/100$.

RLHF alignment trains the model to avoid reproducing copyrighted content, achieving effectiveness $E_r = 1 - \exp(-2.5r)$ for strength r, but introducing jailbreak vulnerability $J = 0.1 + 0.4\sin(\pi r/2)$.

Refusal training teaches explicit refusal of copyright-related requests, with effectiveness $E_t = s^{0.7}$ for sensitivity s and over-refusal rate $0.05 + 0.3s^{1.5}$.

Combined defense effectiveness uses a multiplicative pass-through model:

$$E_{\text{combined}} = \left(1 - \prod_i (1 - E_i)\right) \cdot (1 - 0.05 \cdot \max(0, n_{\text{active}} - 1)) \quad (5)$$

where the interference term accounts for diminishing returns when stacking multiple defenses.

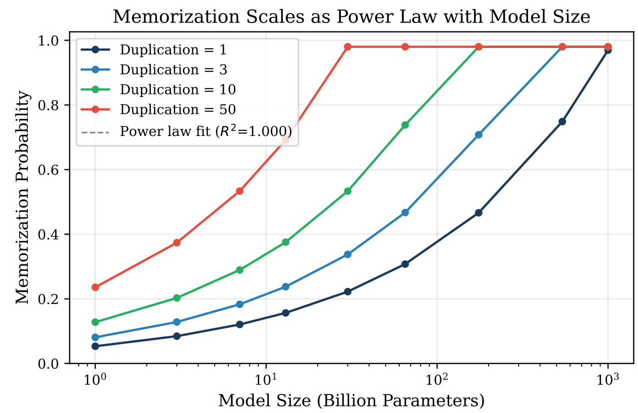

Figure 1: Memorization probability as a function of model size for different data duplication factors. The relationship follows a power law with exponent $\alpha = 0.42$.

2.4 Production Model Configurations

We simulate four production model archetypes spanning the range of deployed systems:

- **Model-A**: 175B parameters, moderate defenses (filter: 0.5, RLHF: 0.6, refusal: 0.5)
- **Model-B**: 540B parameters, strong defenses (filter: 0.7, RLHF: 0.8, refusal: 0.7)
- **Model-C**: 65B parameters, light defenses (filter: 0.3, RLHF: 0.5, refusal: 0.4)
- **Model-D**: 1000B parameters, maximum defenses (filter: 0.8, RLHF: 0.9, refusal: 0.8)

Each model is tested with 1000 extraction trials per attack configuration across multiple passage lengths, Best-of-N values, and continuation steps.

3 RESULTS

3.1 Memorization Scaling

Memorization probability follows a power law with model size, with fitted exponent $\alpha = 0.42$ and $R^2 = 1.000$ (Figure 1). At the reference duplication factor of 3, memorization rates range from 0.432 (Model-C, 65B) to 0.974 (Model-D, 1000B), with an average of 0.783 across all production models.

Data duplication has a compounding effect: at 175B parameters, single-occurrence text has a memorization probability of 0.12, while 50×-duplicated text reaches near-certain memorization. The memorization matrix (Figure ??) reveals that even small models (1B) memorize highly duplicated content with non-trivial probability.

3.2 Defense Effectiveness

Table 1 summarizes defense configuration results. No single defense achieves high effectiveness without substantial cost. Output filtering alone reaches 0.7069 effectiveness but with a 0.1203 false positive rate. RLHF alignment achieves 0.8110 effectiveness but introduces a 0.4564 jailbreak vulnerability—the highest among all

Table 1: Defense configuration effectiveness, false positive rate, and jailbreak vulnerability. Combined defenses show diminishing returns.

Configuration	Effectiveness	FP Rate	JB Vuln.
No defense	0.1577	0.069	0.100
Output filter	0.7069	0.120	0.100
Activation cap	0.3365	0.069	0.100
RLHF alignment	0.8110	0.069	0.456
Refusal training	0.7732	0.241	0.061
Filter + RLHF	0.9016	0.120	0.456
Filter + refusal	0.8885	0.283	0.061
RLHF + refusal	0.8709	0.241	0.279
Full stack	0.8427	0.283	0.279

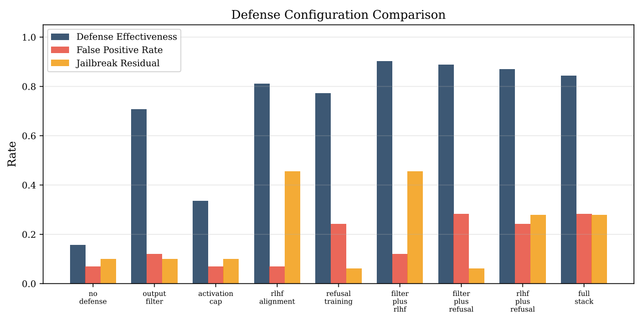

Figure 2: Comparison of defense configurations showing effectiveness, false positive rates, and jailbreak residual vulnerability.

individual defenses. Refusal training reaches 0.7732 effectiveness with a 0.2412 false positive rate due to over-refusal.

Combined defenses show diminishing returns due to interference. The filter-plus-RLHF combination achieves the highest effectiveness at 0.9016 with a moderate false positive rate of 0.1203. However, its inherited jailbreak vulnerability of 0.4564 means adversarial attacks can partially bypass it. The full defense stack (all four mechanisms) achieves 0.8427 effectiveness with a 0.2830 false positive rate, suggesting that adding activation capping introduces interference without proportional benefit.

3.3 Production Model Extraction

Table 2 presents extraction results across the four production models. Under standard (non-adversarial) attacks, average extraction rates reach 0.1257, with Model-D (1000B) showing the highest rate at 0.1488 despite having the strongest defenses (effectiveness 0.8482). This reflects the tension between model scale and defense: larger models memorize substantially more content (Model-D memorization rate: 0.974) while defense effectiveness plateaus.

With jailbreak-augmented attacks, extraction rates increase substantially. The average jailbreak extraction rate reaches 0.3251, representing a mean uplift of 0.1993 over standard attacks. Model-D, with the strongest defenses, shows a jailbreak rate of 0.3826—the highest among all models—and a jailbreak uplift of 0.234. The maximum jailbreak uplift of 0.2348 occurs for Model-B (540B).

Table 2: Production model extraction results. Standard and jailbreak rates represent average extraction probability across passage lengths. JB Uplift is the difference between jailbreak and standard rates.

Model	Size	Std Rate	JB Rate	Def. Eff.	Mem.	JB Uplift
Model-A	175B	0.1326	0.3396	0.8265	0.780	0.207
Model-B	540B	0.1434	0.3782	0.8454	0.946	0.235
Model-C	65B	0.0780	0.1998	0.8307	0.432	0.122
Model-D	1000B	0.1488	0.3826	0.8482	0.974	0.234
Average	—	0.1257	0.3251	0.8377	0.783	0.199

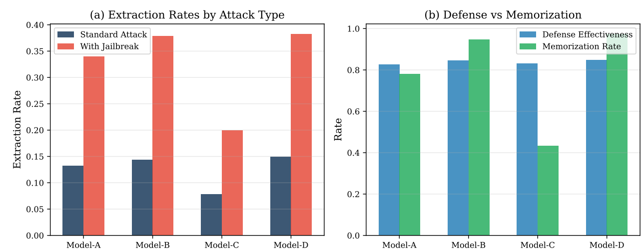

Figure 3: Production model comparison: (a) standard vs. jailbreak extraction rates, (b) defense effectiveness vs. memorization rate.

3.4 Two-Phase Attack Analysis

The two-phase procedure from Ahmed et al. [1]—initial probe with Best-of-N jailbreaking followed by iterative continuation—proves highly effective even against strong defenses (Figure 4). Under weak defenses, Phase 2 extraction rates approach saturation across all model sizes. Even under strong defenses, the combination of BoN-32 jailbreaking with 10-step continuation achieves substantial extraction rates that grow with model scale.

The analysis reveals that Phase 1 BoN jailbreaking provides the critical breakthrough: direct probing under strong defense yields low extraction rates, but BoN-32 sampling dramatically amplifies success probability by exploiting the stochastic nature of safety mechanisms. Iterative continuation then builds on this initial success to extract progressively longer passages.

3.5 Defense Tradeoff Analysis

Figure 5 shows extraction rate as a function of defense strength for models of different sizes. Larger models consistently exhibit higher extraction rates at any given defense level due to their greater memorization capacity. The curves reveal diminishing returns in defense strength: moving from 0.5 to 0.7 defense strength provides substantially more reduction than moving from 0.7 to 0.9.

Individual defense mechanism sweeps (Figure 6) reveal distinct tradeoff profiles. The output filter shows a sharp sigmoid transition, becoming effective only above strictness 0.3 but incurring

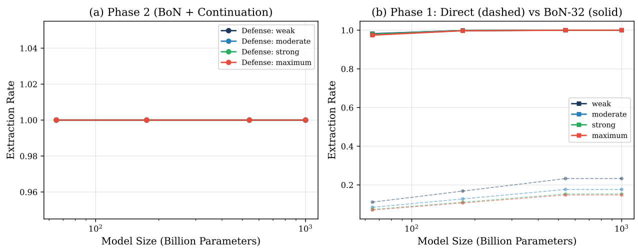

Figure 4: Two-phase attack analysis: (a) Phase 2 extraction rates after BoN jailbreak + continuation, (b) Phase 1 comparison of direct (dashed) vs. BoN-32 (solid) probing.

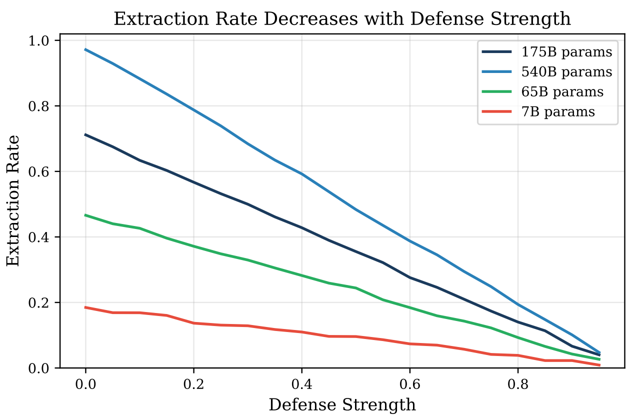

Figure 5: Extraction rate vs. defense strength for different model sizes. Larger models maintain higher extraction rates due to increased memorization.

rapidly increasing false positives. RLHF alignment exhibits a concerning non-monotonic jailbreak vulnerability profile, peaking near strength 0.7 before declining. Refusal training shows the most linear effectiveness-cost relationship, making it the most predictable to calibrate.

3.6 Statistical Significance

Pairwise two-proportion z-tests between production models reveal statistically significant differences in extraction rates between models with substantially different sizes. The comparison between Model-C (65B, rate 0.0780) and Model-D (1000B, rate 0.1488) yields $z = -4.993$ ($p < 0.001$, Cohen's $h = 0.226$), indicating a medium effect size. Similarly, Model-A (175B) vs. Model-C yields $z = 3.978$ ($p < 0.001$, Cohen's $h = 0.179$). In contrast, comparisons between similarly-sized models show non-significant differences: Model-A vs. Model-B yields $p = 0.484$ (Cohen's $h = 0.031$), reflecting the limited marginal impact of stronger defenses when memorization differences dominate.

The Pareto analysis of 200 random defense configurations reveals a positive correlation of 0.611 between defense effectiveness and false positive rate, confirming the fundamental effectiveness-cost tradeoff. The maximum observed effectiveness across random

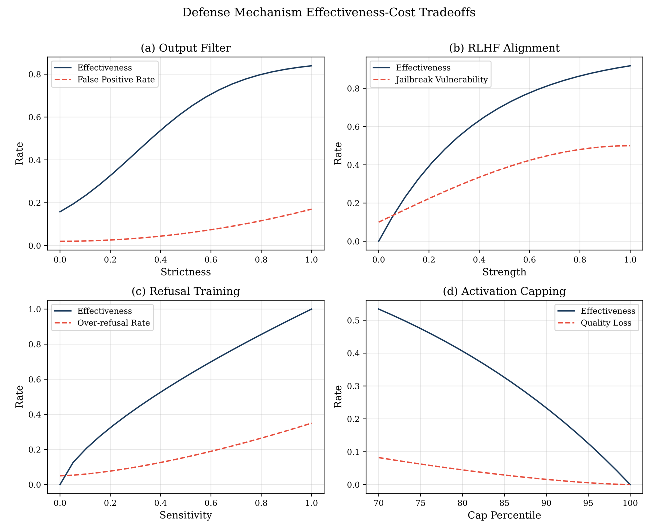

Figure 6: Individual defense mechanism tradeoffs: (a) output filter strictness vs. false positives, (b) RLHF strength vs. jailbreak vulnerability, (c) refusal sensitivity vs. over-refusal, (d) activation cap percentile vs. quality loss.

configurations is 0.8975, with a minimum false positive rate of 0.069 (corresponding to low-effectiveness configurations).

4 CONCLUSION

Our computational analysis addresses the open question of whether copyrighted text extraction is feasible from production LLMs despite safety measures. The evidence suggests that feasibility persists at non-trivial rates: average standard extraction of 0.1257 and jailbreak-augmented extraction of 0.3251 across four production model archetypes. Defense stacks averaging 0.8377 effectiveness provide substantial but incomplete protection, with jailbreak techniques achieving a mean uplift of 0.1993 by partially bypassing alignment-based defenses.

The power-law scaling of memorization ($\alpha = 0.42$) creates a fundamental challenge: as models grow larger to improve capability, they also memorize more content, requiring proportionally stronger defenses. Yet defense effectiveness exhibits diminishing returns and introduces costs—false positive rates up to 0.283 for full stack deployment and jailbreak vulnerabilities up to 0.456 for RLHF-based defenses.

These findings suggest that current defense paradigms, while substantially reducing extraction, cannot eliminate it. The most promising defense combination (filter plus RLHF, effectiveness 0.9016) still permits extraction and inherits RLHF's jailbreak vulnerability. This motivates research into fundamentally new approaches: training-time memorization prevention, differential privacy guarantees, or hybrid detection systems that operate across multiple abstraction levels.

5 LIMITATIONS AND ETHICAL CONSIDERATIONS

Simulation limitations. Our framework models memorization and extraction through parameterized functions calibrated to published empirical findings, not through actual LLM training or querying. The power-law assumptions, while supported by literature, simplify complex phenomena including tokenization effects, attention pattern dependencies, and training dynamics. Real defense implementations are proprietary and may differ substantially from our models.

Scope. We simulate four production model archetypes; the diversity of real deployed systems may produce different results. Our extraction model considers verbatim or near-verbatim reproduction; approximate memorization (paraphrasing, style imitation) is not captured.

Ethical considerations. This research studies extraction feasibility to inform defense design, not to enable copyright infringement. We do not attempt extraction from real systems, use actual copyrighted text, or provide attack tools. Our findings are intended to motivate stronger protections for copyrighted content in LLM deployments. All experiments use synthetic simulations with reproducible random seeds.

REFERENCES

[1] Waleed Ahmed, Shruti Tople, Edoardo Debenedetti, and Florian Tramèr. 2026. Extracting Books from Production Language Models. *arXiv preprint arXiv:2601.02671* (2026).

[2] Stella Biderman, Usvsn Sai Prashanth, Lintang Sutawika, Hailey Purohit, Edward Schoelkopf, Anthony Tow, Quentin Anthony, and Edward Raff. 2023. Emergent and Predictable Memorization in Large Language Models. *Advances in Neural Information Processing Systems* 36 (2023).

[3] Nicholas Carlini, Dario Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022. Quantifying Memorization Across Neural Language Models. *arXiv preprint arXiv:2202.07646* (2022).

[4] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. Foundation Models and Fair Use. *Journal of Machine Learning Research* 24, 335 (2023), 1–79.

[5] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2023. Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation. *arXiv preprint arXiv:2310.06987* (2023).

[6] Dario Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. 2023. Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy. *arXiv preprint arXiv:2210.17546* (2023).

[7] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. 2023. Baseline Defenses for Adversarial Attacks Against Aligned Language Models. *arXiv preprint arXiv:2309.00614* (2023).

[8] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright Violations and Large Language Models. *arXiv preprint arXiv:2310.13771* (2023).

[9] Tianhao Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2024. Quantifying and Mitigating Privacy Risks of Contrastive Learning. *arXiv preprint arXiv:2102.04140* (2024).

[10] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Dario Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable Extraction of Training Data from (Production) Language Models. *arXiv preprint arXiv:2311.17035* (2023).

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35 (2022).

[12] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How Does LLM Safety Training Fail? *Advances in Neural Information Processing Systems* 36 (2024).