Relative Energy Barriers for Copyright-Aware Language Model Adaptation

Wei Sun[1,2] **Yuhang Zhang**[1,2] **Zhicheng Liang**[1,2]
Xianda Wang[3] **Jiaxuan Chen**[1,2] **Fangxin Wang**[1,2*]

[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
[2]Shenzhen Future Network of Intelligence Institute, Shenzhen, China
[3]School of Computing Science, Simon Fraser University, Canada

{weisun2,yuhangzhang,zhichengliang1,jiaxuanchen}@link.cuhk.edu.cn
xianda_wang@sfu.ca wangfangxin@cuhk.edu.cn

Abstract

Adapting language models on long-form corpora can improve domain behavior while also making protected passages easier to reproduce verbatim from short prefixes. We study this tension through *relative energy gain*, the token-level log-density ratio between an adapted model and its pre-adaptation anchor on an exact protected continuation. This quantity separates ordinary predictability from update-induced copying pressure and accumulates over suffixes as a likelihood-ratio advantage for extractable memorization. Building on this view, we introduce Energy-Gated Copyright Regularization (EGCR), a training objective that augments maximum likelihood with a soft relative-energy budget on protected tokens. EGCR uses a differentiable gate combining local persistence of relative gain with anchor surprisal, so regularization is concentrated on distinctive spans whose probabilities rise in a memorization-like way while ordinary next-token supervision is retained elsewhere. Across BookMIA adaptation stress tests, training-prefix extraction, CopyBench benchmark evaluation, and gate ablations, EGCR reduces exact-match and overlap-based reproduction while preserving held-out language-modeling retention, BookMIA-prefix generation quality, and book-related QA utility. Token-level visualizations further show that the gate focuses on a small set of expressive positions associated with verbatim recovery. These results suggest that relative-energy barriers offer an effective and inspectable mechanism for copyright-aware language model adaptation.

1 Introduction

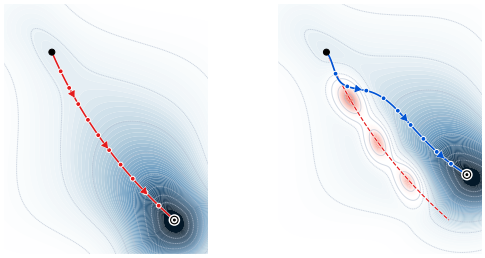
Large language models (LLMs) derive much of their capability from long-form, high-quality text such as books, technical articles, manuals, code, and other curated sources [14, 2, 12], but these documents can also contain expression that is copyright-sensitive or otherwise designated for controlled use. The resulting challenge is not simply whether a model has seen a document, but whether training has changed the conditional distribution so that the protected continuation becomes unusually easy to recover from its prefix. This is the core of extractable verbatim memorization: models can emit training examples under benign or adversarial prompting [4, 5, 34], with risk amplified by scale, duplication, and available prefix context [6, 24, 26, 1]. Yet verbatim recall is entangled with useful learning, because likelihood updates that improve adaptation also lower the energy of observed continuations [20, 13]. Copyright-aware training therefore needs a selective objective that preserves

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

generalizable signal while discouraging the relative suffix likelihood advantage that can support exact reproduction.

Figure 1 illustrates the central intuition. Ordinary likelihood adaptation can turn an observed protected continuation into a favored route through the model's distribution. The desired intervention is to raise a selective barrier on that route while keeping nearby non-verbatim continuations accessible.



(a) Likelihood adaptation **(b)** With a soft barrier

Figure 1: Soft relative-energy barrier intuition. Darker contours mark easier continuations. (a) Likelihood adaptation lowers the protected prefix-to-verbatim route into a copying basin. (b) EGCR raises a selective barrier on that route while keeping non-verbatim alternatives accessible.

An energy view gives a natural way to state this distinction [25, 3]. Low energy corresponds to a continuation that the model can follow easily, so likelihood training can be seen as reshaping the landscape of possible continuations. However, low energy is not a copyright signal. Common syntax, boilerplate, and public facts should remain easy to predict. The useful signal is relative: whether the copyright-sensitive update has made the exact protected continuation newly easier than it was before adaptation. This comparison separates ordinary fluency from update-induced copying pressure.

The energy formulation is also sequence-aware. Verbatim extraction arises when many local preferences align into a stable path from a short prefix to a protected suffix. Measuring energy change along this path connects a token-level training signal to passage-level extraction behavior, echoing sequence-level energy functions for controlled text generation [3, 36]. The result is calibrated to the pre-adaptation distribution, targets newly recoverable routes, and yields inspectable traces of copying pressure.

Within the broader landscape of copyright risk, this paper focuses on extractable verbatim continuation: a measurable behavior in which a protected suffix is regenerated from a prefix. This perspective leads to Energy-Gated Copyright Regularization (EGCR), a training-time mechanism for placing selective barriers during adaptation. Rather than removing protected examples or suppressing all high-probability continuations, EGCR preserves the ordinary likelihood pathway where it supports generalization and adds pressure only when optimization creates a memorization-like route through distinctive protected text. The goal is not to erase the useful linguistic and factual signal in protected corpora, but to prevent adaptation from turning that signal into extractable verbatim continuations.

Our contributions are as follows: **I.** We identify relative continuation energy gain, $\log p_\theta(x_t \mid x_{<t}) - \log p_0(x_t \mid x_{<t})$, as a one-sided risk signal for adaptation-induced verbatim memorization, and show that it additively determines the likelihood-ratio advantage of protected suffixes. **II.** We introduce EGCR, a soft-budget adaptation objective that penalizes excess positive relative gain on protected targets while retaining ordinary next-token supervision. **III.** We propose a dual gate combining local persistence and anchor surprisal, which focuses regularization on distinctive spans that form recoverable copying paths. **IV.** We evaluate EGCR under long-epoch adaptation stress tests with training-prefix extraction, overlap metrics, CopyBench literal-copying evaluation, and token-level diagnostics.

2 Related Work

Memorization and extraction in language models. Training-data memorization is a standard LLM safety and privacy failure mode. Extraction attacks show that models can emit training examples under suitable prompts [4, 5], with risk shaped by capacity, duplication, decoding, and attacker prefix context [6, 24, 26, 1, 40]. For long-form copyrighted works, risk is sequence-level: many conditional predictions can align into a recoverable suffix. Counterfactual analyses separate generic text from examples whose presence changes predictions [43], and latent memorization that reappears under perturbation or relearning [13, 19] motivates objectives that shape memory during adaptation, not only after generation.

Mitigating memorization during and after training. Existing mitigations reduce exposure, alter supervision, or intervene after training. Deduplication reduces a major memorization driver [26, 24], and general regularizers improve generalization, but neither controls which protected continuations become newly recoverable during adaptation [40, 20]. Goldfish Loss modifies the supervised signal by dropping a deterministic subset of target tokens while keeping full input context [16]. Auditing tools improve dataset visibility [30], inference-time filters block exact substrings only after learning has occurred [21], and post-hoc unlearning faces tradeoffs among forgetting, utility, scalability, and robustness to relearning [38, 7, 19]. EGCR instead provides adaptation-time selective control: it retains ordinary supervision while targeting the relative energy gain created during the update.

Copyright-specific objectives and evaluation. Copyright risk differs from generic privacy because practical concerns often center on designated expression, while broader analysis may also consider non-literal similarity in characters, events, style, or expressive structure. Prior work studies plagiarism-like reuse [27], CopyBench literal and non-literal reproduction [9], and memorized book extraction from open-weight models [11]. Other work formalizes copyright protection for generative models [41], proposes copyright-specific objectives [10], or introduces copyright traps [32]. We focus on literal reproduction and develop a training objective for extractable verbatim continuation.

3 Copyright memorization as budgeted energy optimization

Let $\mathcal{D}_{\mathrm{copy}}$ be the protected adaptation corpus. Each document $x = (x_1, \ldots, x_T)$ is treated as a sequence of protected target positions; if only a subspan of a document is protected, the same notation applies after restricting to that span. We write

$$\mathcal{M}_{\mathrm{copy}} = \{(x, t) : x \in \mathcal{D}_{\mathrm{copy}},\ 1 \le t \le |x|\}. \tag{1}$$

Let p_θ denote the model being optimized and p_0 denote the same model before copyright-sensitive adaptation. The anchor p_0 is used only to score the same protected continuations under the pre-update distribution; it is not a teacher distribution over the full vocabulary. The anchor may be any checkpoint immediately preceding this stage, including an instruction-tuned or aligned model; EGCR controls only additional gains introduced during the designated adaptation.

For valid target positions, ordinary maximum likelihood minimizes

$$\mathcal{L}_{\mathrm{copy}}(\theta) = -\frac{1}{|\mathcal{M}_{\mathrm{copy}}|} \sum_{(x,t)\in\mathcal{M}_{\mathrm{copy}}} \log p_\theta(x_t \mid x_{<t}). \tag{2}$$

This objective lowers the energy of protected continuations. The risk is not that a single target token becomes easier, but that many local reductions align into a low-energy path through an exact protected suffix. We therefore define a copyright-risk functional over protected positions,

$$\mathcal{R}_{\mathrm{copy}}(\theta; g) = \frac{1}{|\mathcal{M}_{\mathrm{copy}}|} \sum_{(x,t)\in\mathcal{M}_{\mathrm{copy}}} g_t\, \phi_\tau(\delta_t), \qquad \phi_\tau(z) = [z - \tau]_+^2, \tag{3}$$

where δ_t is the relative energy gain introduced in Section 4, τ is a margin, and $g_t \in [0, 1]$ is a soft selector for positions that carry copy risk. Section 5 instantiates g_t as a detached, model-dependent selector. The hinge-squared barrier ignores negative or sub-margin drift and penalizes smooth excess relative gain, with gradients flowing through the violation term only.

The adaptation objective is then written as a constrained risk minimization problem:

$$\min_\theta \mathcal{L}_{\mathrm{copy}}(\theta) \quad \text{s.t.} \quad \mathcal{R}_{\mathrm{copy}}(\theta; g) \le \varepsilon. \tag{4}$$

The budget ε bounds the average gated violation over the protected corpus. The constraint still permits likelihood learning from $\mathcal{D}_{\mathrm{copy}}$, but discourages update-induced advantages that make exact suffixes newly recoverable; it is a tractable corpus-level control rather than a uniform certificate for every continuation.

4 Relative energy gain for extractable memorization

We now make the relative signal explicit. A causal language model assigns next-token energy $E_\theta(x_t \mid x_{<t}) = -\log p_\theta(x_t \mid x_{<t})$. Because common syntax, boilerplate, and public facts should

remain predictable, absolute energy is not by itself a copyright signal. The relevant quantity is the *additional* energy reduction created by the copyright-sensitive update, measured by the conditional log-density ratio

$$\delta_t = \log p_\theta(x_t \mid x_{<t}) - \log p_0(x_t \mid x_{<t}) = E_0(x_t \mid x_{<t}) - E_\theta(x_t \mid x_{<t}). \tag{5}$$

Positive δ_t means that the update has made the exact token easier than before; negative values mean it has become less preferred. This normalization separates generic predictability from update-induced copying pressure: a token may be likely because it is common, or unlikely in absolute terms yet suspicious because adaptation sharply increased its probability.

This token-level quantity matters because verbatim extraction is a sequence event.

Lemma 1 (Additivity of relative energy gain)**.** *For any protected suffix $x_{a:b}$ under prefix $x_{<a}$, the relative log-likelihood of the whole suffix decomposes into token-level relative energy gains:*

$$\log \frac{p_\theta(x_{a:b} \mid x_{<a})}{p_0(x_{a:b} \mid x_{<a})} = \sum_{t=a}^{b} \delta_t. \tag{6}$$

Proof. Both p_θ and p_0 are causal language models, so $p_\theta(x_{a:b} \mid x_{<a}) = \prod_{t=a}^{b} p_\theta(x_t \mid x_{<t})$ and similarly for p_0. Taking the log of the ratio turns the product of conditional ratios into the sum in Eq. (6). □

Definition 1 (Relative suffix likelihood advantage)**.** *For a protected suffix $x_{a:b}$ under prefix $x_{<a}$, define*

$$A_\theta(a, b) = \log \frac{p_\theta(x_{a:b} \mid x_{<a})}{p_0(x_{a:b} \mid x_{<a})} = \sum_{t=a}^{b} \delta_t. \tag{7}$$

The statistic $A_\theta(a, b)$ measures the log likelihood-ratio advantage of the exact suffix under the adapted model. Large positive values indicate that the update has created a lower-energy path through the protected continuation, independently of the absolute probability assigned by either model.

Proposition 1. *For any protected suffix $x_{a:b}$ and prefix $x_{<a}$, bounding the cumulative relative energy gain $A_\theta(a, b) \le B$ bounds the multiplicative increase in suffix probability by*

$$\frac{p_\theta(x_{a:b} \mid x_{<a})}{p_0(x_{a:b} \mid x_{<a})} \le \exp(B). \tag{8}$$

In particular, if $\delta_t \le \tau$ for every $t \in [a, b]$, then the suffix likelihood ratio over a suffix of length $n = b - a + 1$ is at most $\exp(n\tau)$.

Proof. The first claim follows by exponentiating Eq. (7). If each token-level gain is at most τ, then $A_\theta(a, b) = \sum_{t=a}^{b} \delta_t \le n\tau$. □

Proposition 2 (From local quadratic excess to a suffix barrier)**.** *Fix a protected suffix $x_{a:b}$ of length $n = b - a + 1$ and suppose its gate values satisfy $g_t \ge \underline{g} > 0$ on the suffix. Let*

$$V_\theta(a, b; g) = \sum_{t=a}^{b} g_t [\delta_t - \tau]_+^2. \tag{9}$$

Then

$$A_\theta(a, b) \le n\tau + \sqrt{\frac{n\, V_\theta(a, b; g)}{\underline{g}}}. \tag{10}$$

Proof. For every t, $\delta_t \le \tau + [\delta_t - \tau]_+$. Summing over the suffix and applying Cauchy–Schwarz gives

$$A_\theta(a, b) \le n\tau + \sum_{t=a}^{b} [\delta_t - \tau]_+ \le n\tau + \sqrt{n \sum_{t=a}^{b} [\delta_t - \tau]_+^2}.$$

Since $g_t \ge \underline{g}$, the unweighted squared excess is at most $V_\theta(a, b; g)/\underline{g}$, which proves the claim. □

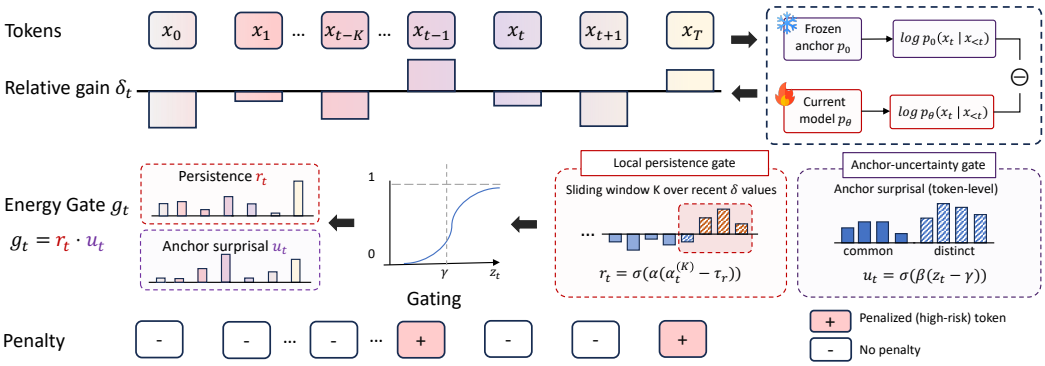


Figure 2: Overview of Energy-Gated Copyright Regularization (EGCR). EGCR compares each protected token under the current model and a frozen anchor to trace update-induced relative energy gain. A detached gate combines local persistence with anchor surprisal, so the penalty concentrates on distinctive spans whose probabilities have increased in a sustained, memorization-like way while leaving low-risk tokens under ordinary likelihood training.

These propositions motivate the barrier view. The margin τ sets tolerated per-token drift, but the $\exp(n\tau)$ bound can be loose for long suffixes. The useful control comes from penalizing excess positive drift on sustained, high-gate spans: Proposition 2 shows that small local quadratic excess rules out arbitrarily large relative suffix advantage while allowing distributional change elsewhere.

5 Energy-gated copyright regularization

Figure 2 summarizes EGCR. The selector g_t in Eq. (3) approximates whether position t participates in a protected continuation whose probability has increased in a memorization-like way. We decompose it into local persistence and anchor-surprisal components,

$$g_t = r_t u_t, \tag{11}$$

where r_t measures persistent relative gain and u_t measures anchor difficulty.

Local persistence gate. Extraction is sequential, so an isolated positive δ_t is less informative than a run of gains. Let $\kappa_0, \ldots, \kappa_{K-1}$ be a causal smoothing kernel with $\kappa_j \geq 0$ and $\sum_j \kappa_j = 1$. We define a detached local relative-likelihood signal

$$a_t^{(K)} = \sum_{j=0}^{K-1} \kappa_j \, \mathrm{stopgrad}(\delta_{t-j}), \qquad r_t = \sigma\left(\alpha(a_t^{(K)} - \tau_r)\right). \tag{12}$$

The uniform kernel $\kappa_j = 1/K$ recovers a moving average. Larger values of r_t correspond to windows in which the protected continuation has gained probability consistently relative to the anchor. The stop-gradient operator prevents the model from reducing the penalty by changing the selector; the penalty gradient is carried by the violation term itself. At the beginning of a protected span, the sum is taken only over valid previous protected positions and the kernel is renormalized, so smoothing windows do not cross document or span boundaries.

Anchor-surprisal gate. The second factor distinguishes generic continuations from distinctive ones. Let $s_t = -\log p_0(x_t \mid x_{<t})$ be the anchor surprisal of the protected token. Using corpus-level statistics over valid protected tokens, define

$$z_t = \frac{s_t - \mu_s}{\sigma_s + \eta}, \qquad u_t = \sigma(\beta(z_t - \gamma)). \tag{13}$$

Here γ controls the surprisal threshold and η is a numerical constant. In our experiments, μ_s and σ_s are precomputed once from the protected adaptation corpus under the frozen anchor. Tokens that were already predictable under p_0 receive smaller u_t, even if they occur in protected documents. Tokens with high anchor surprisal receive larger weight, but only become strongly penalized when paired with persistent relative energy gain through r_t.

The resulting copyright regularizer is

$$\mathcal{L}_{\mathrm{EGCR}}(\theta) = \frac{1}{|\mathcal{M}_{\mathrm{copy}}|} \sum_{(x,t)\in\mathcal{M}_{\mathrm{copy}}} \mathrm{stopgrad}(r_t u_t)\,[\delta_t - \tau]_+^2, \tag{14}$$

and EGCR optimizes the Lagrangian relaxation

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{copy}}(\theta) + \lambda \mathcal{L}_{\mathrm{EGCR}}(\theta). \tag{15}$$

For active violations, the gradient of the copyright term takes the form

$$\nabla_\theta \mathcal{L}_{\mathrm{EGCR}} = \frac{2}{|\mathcal{M}_{\mathrm{copy}}|} \sum_{(x,t)\in\mathcal{M}_{\mathrm{copy}}} \mathrm{stopgrad}(r_t u_t)[\delta_t - \tau]_+ \nabla_\theta \log p_\theta(x_t \mid x_{<t}), \tag{16}$$

where the anchor term has zero gradient. Equivalently, for one active token and a fixed detached gate, the derivative of

$$-\log p_\theta(x_t \mid x_{<t}) + \lambda g_t [\delta_t - \tau]_+^2 \tag{17}$$

with respect to $\log p_\theta(x_t \mid x_{<t})$ is

$$-1 + 2\lambda g_t(\delta_t - \tau). \tag{18}$$

Thus exceeding the margin does not by itself guarantee that the total update will push the token probability downward. Above the margin, EGCR reduces the effective likelihood incentive and, for sufficiently large violations or dual weights, reverses it. Below the margin, training is identical to ordinary likelihood optimization on that token.

The multiplier λ can be treated as a fixed hyperparameter or updated as a dual variable for a target budget ε:

$$\lambda_{m+1} = \left[\lambda_m + \rho\left(\mathcal{R}_{\mathrm{copy}}(\theta_m; g) - \varepsilon\right)\right]_+, \tag{19}$$

with dual step size ρ. This form makes the tradeoff explicit: the model uses ordinary likelihood learning unless the measured relative energy gain risk exceeds the allowed budget.

Implementation. EGCR requires only quantities already available in next-token training plus anchor log probabilities from p_0. For a fixed protected corpus, the anchor log probabilities and surprisals can be precomputed once. For online or mixed adaptation, they can be obtained from a frozen copy of the starting model. During training, protected examples receive the additional loss in Eq. (14). If the practitioner also includes auxiliary non-copyright data for utility retention, those examples simply use the ordinary language-modeling loss and do not enter $\mathcal{R}_{\mathrm{copy}}$. Thus EGCR is compatible with standard fine-tuning, continued pretraining, parameter-efficient adaptation, and mixed-data retention schedules. Unlike Goldfish Loss, it keeps full target-token supervision; unlike Copyright Regression, it directly penalizes the update-induced likelihood-ratio advantage of exact continuations rather than applying a corpus-level copyright term uniformly.

6 Experimental Settings

Models and Implementation. We evaluate Llama3-8B, Llama3.2-3B, Qwen3-4B, and Qwen3-8B [15, 33, 42] in a copyright-sensitive adaptation setting. The goal is to test whether a model can learn from a designated protected corpus without acquiring a high-probability path for verbatim reproduction. All methods are implemented with LoRA adaptation [18] on the same protected adaptation corpus and evaluated under the same decoding setup. Unless otherwise stated, reported values are averages over five independent runs. We also measure EGCR's per-epoch training time and additional transient memory per GPU relative to LoRA-SFT under matched settings; details are given in Appendix B.4.

Baselines. We compare EGCR with LoRA-SFT, LoRA-SFT with AdamW weight decay, Goldfish Loss (GL) [16], and Copyright Regression (CR) [10]. LoRA-SFT optimizes the ordinary next-token language-modeling loss on the designated protected corpus; the weight-decay variant applies $w_d = 0.1$ to trainable LoRA parameters. GL masks a content-dependent subset of target tokens, while CR adds an explicit copyright-data term. All methods use the same base models, BookMIA corpus, and evaluation prompts. We use a label-scarce protected-adaptation protocol rather than assuming a clean copyright/non-copyright split in the adaptation data; Appendix A discusses this choice.

Table 1: Main BookMIA results across training checkpoints. Reproduction metrics are ROUGE-L (R-L)↓ and rescaled BERTScore (BS)↓; retention and quality metrics are WikiText-2 perplexity (PPL)↓ and BookMIA-continuation MAUVE (MV)↑. Values are mean ± standard deviation.

Model	Method	Epoch 3				Epoch 10				Epoch 30			
		R-L	BS	PPL	MV	R-L	BS	PPL	MV	R-L	BS	PPL	MV
Llama3.2 3B	LoRA-SFT	0.18± 0.01	0.10± 0.01	18.08± 0.32	0.88± 0.01	0.27± 0.02	0.19± 0.02	28.86± 1.12	0.89± 0.01	0.72± 0.03	0.67± 0.03	142.09± 6.84	0.91± 0.01
	GL	0.17± 0.01	0.08± 0.01	18.06± 0.28	**0.89**± 0.01	0.20± 0.01	0.10± 0.01	33.16± 1.45	0.87± 0.02	0.24± 0.02	0.14± 0.01	193.41± 8.72	0.91± 0.01
	CR	0.18± 0.01	0.09± 0.01	18.13± 0.41	0.86± 0.02	0.25± 0.02	0.17± 0.02	24.99± 0.96	0.87± 0.02	0.35± 0.03	0.36± 0.03	123.74± 5.91	0.91± 0.01
	EGCR	**0.16**± 0.01	**0.07**± 0.01	**15.67**± 0.30	0.88± 0.01	**0.18**± 0.01	**0.09**± 0.01	**17.60**± 0.44	**0.90**± 0.01	**0.20**± 0.01	**0.11**± 0.01	**21.27**± 0.77	**0.93**± 0.01
Llama3 8B	LoRA-SFT	0.39± 0.02	0.32± 0.02	14.03± 0.25	**0.92**± 0.01	0.63± 0.03	0.59± 0.03	24.14± 0.88	0.93± 0.01	0.93± 0.02	0.92± 0.02	136.39± 7.36	0.93± 0.01
	GL	0.30± 0.02	0.22± 0.02	14.01± 0.22	0.89± 0.02	0.32± 0.02	0.24± 0.02	27.49± 1.06	0.90± 0.02	0.34± 0.02	0.27± 0.02	143.70± 6.58	0.92± 0.01
	CR	0.32± 0.02	0.26± 0.02	**11.08**± 0.21	0.85± 0.02	0.44± 0.03	0.49± 0.03	22.34± 0.74	0.93± 0.01	0.63± 0.04	0.63± 0.03	93.36± 4.81	**0.94**± 0.01
	EGCR	**0.21**± 0.01	**0.12**± 0.01	12.26± 0.27	**0.92**± 0.01	**0.24**± 0.02	**0.14**± 0.01	**13.71**± 0.38	**0.94**± 0.01	**0.26**± 0.02	**0.17**± 0.01	**15.70**± 0.52	0.93± 0.01
Qwen3 4B	LoRA-SFT	0.25± 0.02	0.17± 0.01	20.17± 0.47	0.83± 0.02	0.30± 0.02	0.22± 0.02	34.34± 1.31	0.90± 0.01	0.68± 0.04	0.66± 0.04	272.20± 12.64	0.89± 0.02
	GL	0.25± 0.02	0.16± 0.01	20.46± 0.51	0.83± 0.02	0.26± 0.02	0.18± 0.02	41.83± 1.83	0.88± 0.02	0.29± 0.02	0.22± 0.02	487.82± 18.91	0.89± 0.02
	CR	0.25± 0.02	0.17± 0.02	**20.06**± 0.44	0.80± 0.02	0.23± 0.02	0.21± 0.02	33.80± 1.27	0.88± 0.02	0.46± 0.03	0.43± 0.03	248.82± 11.37	0.87± 0.02
	EGCR	**0.16**± 0.01	**0.07**± 0.01	22.70± 0.58	**0.91**± 0.01	**0.17**± 0.01	**0.07**± 0.01	**23.47**± 0.64	**0.91**± 0.01	**0.21**± 0.02	**0.11**± 0.01	**27.96**± 0.83	**0.91**± 0.01
Qwen3 8B	LoRA-SFT	0.26± 0.02	0.18± 0.02	16.66± 0.36	0.87± 0.02	0.32± 0.02	0.25± 0.02	25.88± 0.92	**0.93**± 0.01	0.72± 0.03	0.70± 0.03	262.09± 10.84	0.93± 0.01
	GL	0.25± 0.02	0.17± 0.01	16.84± 0.34	0.85± 0.02	0.26± 0.02	0.18± 0.02	31.87± 1.18	0.85± 0.02	0.30± 0.02	0.23± 0.02	376.64± 15.73	0.88± 0.02
	CR	0.25± 0.02	0.17± 0.01	**16.60**± 0.32	**0.88**± 0.01	0.31± 0.02	0.23± 0.02	28.61± 1.04	0.87± 0.02	0.54± 0.03	0.48± 0.03	184.05± 8.22	**0.94**± 0.01
	EGCR	**0.15**± 0.01	**0.06**± 0.01	16.82± 0.39	**0.88**± 0.01	**0.17**± 0.01	**0.08**± 0.01	**17.98**± 0.48	**0.93**± 0.01	**0.19**± 0.01	**0.09**± 0.01	**21.66**± 0.69	0.93± 0.01

Data and Evaluation. The primary protected adaptation corpus is BookMIA [39, 8], which we designate as protected under our experimental protocol for adaptation and training-prefix copying evaluation. Copying risk is measured on BookMIA using exact-match rates [5, 34], ROUGE-L [28], and BERTScore [44] against protected training continuations. This evaluation asks whether the adapted model can reproduce training text verbatim from an in-corpus prefix. We use this in-corpus prefix protocol as an extraction stress test for adaptation-induced memorization, not as a claim that the benchmark covers all forms of copyright-sensitive adaptation or legal infringement. We separate two non-copying notions of utility: retention of general language-modeling ability is measured by WikiText-2 perplexity [31, 22], while generation quality under BookMIA prefixes is measured by MAUVE [35] between sampled continuations and the corresponding BookMIA continuation set. We also evaluate on the Literal Copying and Fact Recall subsets of CopyBench [9]. Results are reported at Epochs 3, 10, and 30, with Epoch 30 serving as a long-epoch memorization stress test.

Full training details, hyperparameters, dataset construction, decoding, and metric definitions are provided in Appendix A.

7 Results

We present three evaluations of EGCR: BookMIA reproduction–retention tradeoffs, training-prefix extraction, and CopyBench benchmark evaluation in Table 1, Figure 3, and Figure 4.

EGCR reduces reproduction while preserving retention and quality. Table 1 shows that EGCR jointly suppresses protected-continuation reproduction, avoids held-out PPL blow-up, and maintains high BookMIA-continuation MAUVE under long protected-data adaptation. The PPL gap is expected: LoRA-SFT overfits the protected-book distribution, GL removes much target supervision, and CR mostly rescales corpus-level fitting in the all-protected setting. EGCR preserves ordinary supervision while blocking sustained anchor-relative gains, so its low PPL and high MAUVE reflect lower copying with retained held-out general-text modeling and without collapsed BookMIA-prefix generations; see Appendix C.

Weight-decay comparison and generation quality. At Epoch 30, LoRA-SFT with weight decay averages 0.68 ROUGE-L, 0.65 BERTScore, 193.22 WikiText-2 PPL, and 0.92 MAUVE across the four backbones, compared with 0.76, 0.74, and 203.19 for LoRA-SFT on the first three metrics. Rubric-based quality scores are also highest for EGCR at Epoch 30 on both WikiText-2 and BookMIA for every backbone. Full per-model results are given in Appendices B.1 and B.2.

EGCR resists training-prefix extraction. Figure 3 reports BookMIA training-prefix extraction in the main Llama3-8B stress setting: given a protected training prefix, the attacker tries to recover the following continuation, and EM@k counts only exact token matches for the first k generated tokens. EGCR keeps extraction near the low-copying baselines across continuation lengths and training epochs, indicating that the energy-gated penalty interrupts recoverable verbatim chains rather than merely reducing soft overlap. Appendix D analyzes the corresponding cross-model extraction results.

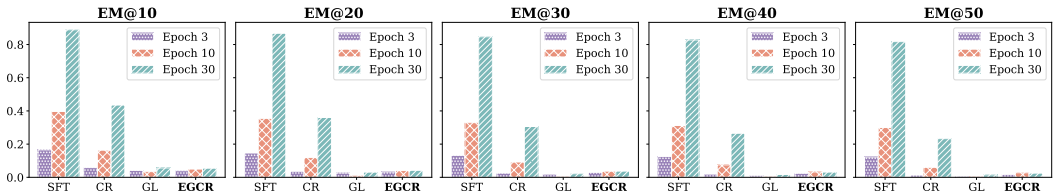


Figure 3: Training-prefix extraction on BookMIA for Llama3-8B. EM@k is the fraction of prompts whose first k generated tokens exactly match the protected reference continuation.

EGCR reduces literal copying while preserving utility on CopyBench. CopyBench complements BookMIA by separating protected-expression copying from question-answering behavior about book content. We report two CopyBench metrics: `rouge_l_ge_0.8_rate`, the probability that a generation has ROUGE-L ≥ 0.8 against the reference, and QA `mean_f1`, the mean F1 score for answer correctness.

Figure 4 shows that EGCR is favorable on both axes. LoRA-SFT accumulates literal copying, and CR only partially slows this growth. GL keeps copying near the base rate, but its QA F1 declines, suggesting weakened book-related learning. EGCR follows the low-copying trajectory without that utility loss: its literal-copy rate stays close to the base model, while its QA F1 stays above the baseline and is strongest among adapted methods at late checkpoints. Additional trajectory analysis and representative input–reference examples are provided in Appendix E. QA F1 measures answer correctness, whereas literal-copy rate measures textual overlap; together they track book-related utility and verbatim reproduction separately.

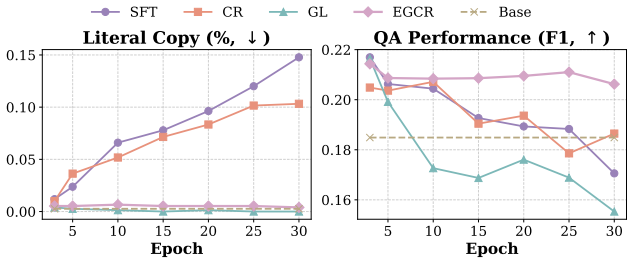


Figure 4: CopyBench benchmark evaluation across adaptation epochs.

8 Analysis

The results above show that EGCR reduces reproduction metrics while preserving general-language retention and BookMIA-prefix generation quality. We now analyze generations, token-level selectivity, and the dual gate.

Qualitative extraction behavior. Figure 5 provides representative generations under the same protected member prefix from *Emma*. We show Llama3-8B at Epoch 3 because this checkpoint is closest to a realistic adaptation scenario: the model has absorbed useful domain signal without long-epoch memorization stress. The example tests whether a method reduces extractable copying without relying on overtraining. Additional qualitative examples are provided in Appendix F.

Prompt	it was known all over Highbury.It was the very event to engage those who talk most, the young and the low; and all the youth and servants in the place were soon in the happiness of frightful news.
Reference	The last night's ball seemed lost in the gipsies.Poor Mr.Woodhouse trembled as he sat, and, as Emma had foreseen, would scarcely be satisfied without their promising never to go beyond the shrubbery again.
SFT	The last night's ball seemed lost in the gipsies. Poor Mr. Woodhouse trembled as he sat, and, as Emma had foreseen, would scarcely be satisfied without their promising never to go beyond the shrubbery again.
CR	The last night's ball seemed lost in the gipsies. Poor Mr. Woodhouse trembled as he sat, and, as Emma had foreseen, would scarcely be satisfied without their promising never to go beyond the shrubbery again.
GL	The last night's ball was not talked of half so much as the calamity of this morning. The second instance of failing health, coming after such a gap against such a character, overwhelmed emotion.
EGCR	It was a great pleasure to Mr. Weston; though he had no hand in it, he was very ready to be concerned, and to tell, with much delight, his own fears and suspicions, with all the particulars of their first discovery in Hartfield.

Figure 5: Qualitative training-prefix extraction on a protected *Emma* passage for Llama3-8B at Epoch 3. Green marks coherent non-copied continuation, red marks copied or strongly overlapped spans, and orange marks degraded output quality. More examples are in Appendix F.

The qualitative behavior mirrors the aggregate extraction results. Standard fine-tuning continues along the protected reference nearly verbatim, and Copyright Regression exhibits the same failure because its corpus-level objective does not directly target the token-by-token continuation advantage

created during adaptation. Goldfish Loss interrupts the exact chain after an initial overlap, but the continuation becomes less stable once local target supervision is removed. EGCR instead diverges early from the reference while remaining fluent and contextually coherent. This pattern matches the design of the objective: EGCR keeps ordinary next-token supervision available, but applies selective pressure when sustained relative gains emerge on distinctive protected tokens.

Token-level selectivity. Figure 6 gives one Llama3-8B qualitative example of the kind of selectivity produced by the gate; additional examples are provided in Appendix G. The highlighted positions occupy only a small fraction of the passage, so the method does not need to modify every protected target token in order to change the reproduction behavior. This contrasts with fixed token dropping: EGCR can keep ordinary supervision on most positions and spend its regularization budget on a narrower set of tokens that are more relevant to verbatim recovery. This more economical intervention is consistent with the tradeoff in Table 1, where EGCR achieves low reproduction scores without the large retention or generation-quality degradation seen in less selective interventions.

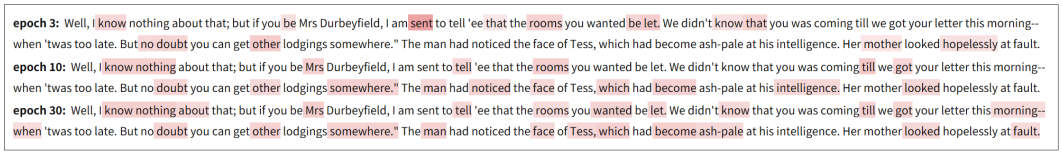


Figure 6: Token-level selectivity on a protected *Tess of the d'Urbervilles* passage for Llama3-8B. Darker highlights denote stronger gate activation. Other models are in Appendix G.

The selected tokens also show where memorization pressure tends to become visible in literary text. The gate is often triggered around distinctive wording, names, dialectal phrasing, and phrase endings such as "Mrs Durbeyfield," "rooms," "somewhere," and "ash-pale." These tokens are not merely syntactic fillers; they help pin down the exact style and continuation of the protected passage. As training proceeds from Epoch 3 to Epoch 30, the highlighting becomes more stable around such expressive choices, suggesting that the gate is responding to the parts of the passage most likely to support verbatim continuation rather than applying uniform pressure across the text.

Gate dynamics. We quantify activation using the fraction of valid protected positions with $g_t = r_t u_t > 0.5$ and the mean gate value. Across all four backbones, the high-gate fraction rises from Epoch 3 to 15 and changes more modestly by Epoch 30, while mean activation continues to increase. Full results are in Appendix B.3.

Ablating the dual gate. We isolate the two factors in the EGCR gate by retraining under the same BookMIA adaptation protocol, optimizer, LoRA configuration, and EGCR hyperparameters, changing only the selector g_t inside Eq. (14). The ablation is reported on the Llama3-8B Epoch 30 checkpoint because this long-exposure setting makes the memorization–retention tradeoff most visible.

Table 2: Gate ablation.

Gate	R-L ↓	BS ↓	PPL ↓	MV ↑
$g_t = 1$	0.27	0.11	375.72	0.79
$g_t = r_t$	0.21	0.09	323.33	0.74
$g_t = u_t$	0.35	0.24	93.89	0.84
$g_t = r_t u_t$	0.26	0.17	15.70	0.93

Table 2 supports the role of the dual gate. Removing the gate or using only the persistence factor applies a much less selective regularization signal: it can reduce surface reproduction, but it also spends the regularization budget on many positions that are not the most useful targets for extractable continuations. The surprisal-only variant is more conservative, but it lacks the sequential check needed to identify recoverable verbatim chains. The product gate gives the intended behavior: it activates when a token is both distinctive under the anchor and part of a persistent relative energy gain, yielding the strongest overall reproduction–retention tradeoff rather than merely the strongest penalty.

Hyperparameter sensitivity. We evaluate one-at-a-time changes to λ, τ, and K on Llama3.2-3B. Their effects become clearer with longer adaptation: at Epoch 30, setting $\lambda = 0.5$ raises ROUGE-L/BERTScore from 0.20/0.11 to 0.27/0.16, while $\tau = 1.0$ raises them to 0.31/0.23; $K = 16$ yields 0.23/0.14. Full results at Epochs 3, 10, and 30 are reported in Table 3.

Table 3: One-at-a-time EGCR hyperparameter sensitivity on Llama3.2-3B. Default uses $\lambda = 1$, $\tau = 0.5$, and $K = 8$; each variant changes only the indicated parameter.

Setting	Epoch 3				Epoch 10				Epoch 30			
	R-L	BS	PPL	MV	R-L	BS	PPL	MV	R-L	BS	PPL	MV
Default	0.16	0.07	15.67	0.88	0.18	0.09	17.60	0.90	0.20	0.11	21.27	0.93
$\lambda = 0.5$	0.17	0.08	15.92	0.86	0.19	0.11	19.16	0.88	0.27	0.16	27.25	0.90
$\lambda = 2.0$	0.16	0.07	15.50	0.87	0.17	0.08	16.62	0.86	0.19	0.11	19.15	0.89
$\tau = 0.25$	0.17	0.06	15.55	0.86	0.18	0.09	17.26	0.86	0.20	0.11	20.93	0.91
$\tau = 1.0$	0.17	0.08	16.86	0.89	0.21	0.11	28.37	0.87	0.31	0.23	33.84	0.93
$K = 4$	0.17	0.08	15.60	0.87	0.18	0.10	17.57	0.87	0.20	0.12	21.72	0.89
$K = 16$	0.17	0.08	15.74	0.85	0.20	0.10	19.69	0.90	0.23	0.14	27.91	0.90

9 Ethical Considerations and Limitations

EGCR is designed to reduce adaptation-induced extractable verbatim continuations from designated protected prefixes. Here, "protected" refers to the designation made by our experimental protocol; it is not a legal determination of the copyright status of any work or of its status in any jurisdiction.

Ethical considerations and dual use. EGCR may support authorized adaptation, but could also be misused to suppress evidence of regurgitation after unauthorized training. Low extraction is not evidence that a work was absent from training, lawfully acquired, or removed from model parameters. EGCR is neither a compliance certificate nor a provenance or unlearning method; it should complement licensing review, auditable training records, and independent extraction audits. It is also complementary to output-time similarity filters and post-hoc unlearning, which address generated outputs and memorization acquired before the designated adaptation stage, respectively.

Empirical scope. Experiments cover English literary text in BookMIA and selected CopyBench tasks, four model families up to 8B parameters, and extraction from exact training prefixes. They do not establish robustness on other domains, languages, larger models, paraphrased or adversarial prompts, or multi-turn attacks. Since the objective targets exact token sequences, it does not directly address paraphrase, style imitation, plot reuse, or broader substantial similarity; preserved QA utility is not a legal assessment of retained information.

Methodological limitations. EGCR only controls gains after the chosen anchor and does not address prior memorization. Its corpus-level objective is not a worst-case guarantee; the suffix bound in Proposition 2 requires a positive gate lower bound. Gate settings require calibration, and the selective signal may miss expressive spans or flag legitimate content. Our utility metrics are incomplete proxies for these effects.

10 Conclusion and Future Work

EGCR is a copyright-sensitive objective that controls memorization with a relative-energy barrier against a pre-adaptation anchor. It identifies tokens with distinctive, persistent likelihood gains, penalizes large gated budget violations, and preserves supervision elsewhere. Across BookMIA, training-prefix extraction, CopyBench evaluation, ablations, and token analyses, EGCR reduces verbatim reproduction while preserving held-out LM retention, BookMIA-prefix quality, and book-related QA utility, showing that it can curb extractable copying without discarding useful learning.

Future work should sharpen gate selectivity, moving from a memorization-oriented training signal toward more interpretable token-level protection. More broadly, EGCR's stable low PPL across epochs suggests that relative-energy gating preserves predictive fluency under repeated adaptation and may be useful beyond copyright: in continual or lifelong learning, similar controls could help models absorb new domain knowledge while preserving general predictive behavior and prior behavior.

Acknowledgments

The work was supported in part by the Guangdong S&T Programme (Grant No. 2024B0101030002), the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, the National Key Research and Development Program of China (Grant No. 2024YFB2907000), the National Natural Science Foundation of China (Grant No. 62293482 and Grant No. 62471423), the Shenzhen Science and Technology Program (Grant No. JCYJ20241202124021028 and Grant No. JCYJ20230807114204010), the Guangdong Talents Program (Grant No. 2024TQ08X346), the Shenzhen Outstanding Talents Training Fund 202002, the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. SYSPG20241211173853027), and the Guangdong Province Radio Science Data Center (Grant No. 2025B1212070001).

References

[1] Stella Biderman, Usvsn Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems*, 2023.

[2] Stella Biderman, Kieran Bicheno, and Leo Gao. Datasheet for the Pile. *arXiv preprint arXiv:2201.07311*, 2022.

[3] Anton Bakhtin, Yuntian Deng, Sam Gross, Myle Ott, Marc'Aurelio Ranzato, and Arthur Szlam. Residual energy-based models for text. *Journal of Machine Learning Research*, 22(40):1–41, 2021.

[4] Nicholas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos, and Dawn Song. The Secret Sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, 2019.

[5] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

[6] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023.

[7] Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. Towards robust and parameter-efficient knowledge unlearning for LLMs. In *International Conference on Learning Representations*, 2025.

[8] Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, Memory: An archaeology of books known to ChatGPT/GPT-4. In *Empirical Methods in Natural Language Processing*, 2023.

[9] Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. CopyBench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. In *Empirical Methods in Natural Language Processing*, 2024.

[10] Timothy Chu, Zhao Song, and Chiwun Yang. How to protect copyright data in optimization of large language models? In *AAAI Conference on Artificial Intelligence*, 2024.

[11] A. Feder Cooper, Mark A. Lemley, Allison Casasola, Ahmed Ahmed, Aaron Gokaslan, Amy B. Cyphert, Christopher De Sa, Daniel E. Ho, and Percy Liang. Extracting memorized pieces of (copyrighted) books from open-weight language models. *arXiv preprint arXiv:2505.12546*, 2025.

[12] Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In *Empirical Methods in Natural Language Processing*, 2021.

[13] Sunny Duan, Mikail Khona, Abhiram Iyer, Rylan Schaeffer, and Ila Rani Fiete. Uncovering latent memories in large language models. In *International Conference on Learning Representations*, 2025.

[14] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, and others. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

[16] Abhimanyu Hans, Yuxin Wen, Neel Jain, John Kirchenbauer, Hamid Kazemi, Prajwal Singhania, Siddharth Singh, Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, and Tom Goldstein. Be like a Goldfish, don't memorize! Mitigating memorization in generative LLMs. In *Advances in Neural Information Processing Systems*, 2024.

[17] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[19] Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? Jogging the memory of unlearned LLMs via benign relearning. In *International Conference on Learning Representations*, 2025.

[20] Jing Huang, Diyi Yang, and Christopher Potts. Demystifying verbatim memorization in large language models. In *Empirical Methods in Natural Language Processing*, 2024.

[21] Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette-Choo, and Nicholas Carlini. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *International Natural Language Generation Conference*, 2023.

[22] Fred Jelinek, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. Perplexity: A measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977.

[23] Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient LLM unlearning framework from logit difference. In *Advances in Neural Information Processing Systems*, 2024.

[24] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, 2022.

[25] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006.

[26] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Association for Computational Linguistics*, 2022.

[27] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. Do language models plagiarize? In *ACM Web Conference*, 2023.

[28] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[30] Marc Marone and Benjamin Van Durme. Data Portraits: Recording foundation model training data. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.

[31] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[32] Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre de Montjoye. Copyright traps for large language models. In *International Conference on Machine Learning*, 2024.

[33] Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`, 2024.

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

[35] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, 2021.

[36] Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. COLD decoding: Energy-based constrained text generation with Langevin dynamics. In *Advances in Neural Information Processing Systems*, 2022.

[37] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

[38] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, 2025.

[39] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

[40] Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*, 2022.

[41] Nikhil Vyas, Sham Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International Conference on Machine Learning*, 2023.

[42] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, and others. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[43] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramer, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems*, 2023.

[44] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

[45] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling*, 2024.

A Experimental details

Training setup. All adaptation experiments use LoRA with rank $r = 32$, scaling parameter $\alpha_{\text{LoRA}} = 64$, and dropout 0.05 [18]. LoRA is applied to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. Base models are loaded and trained in `torch.bfloat16`. We optimize with AdamW [29] using learning rate 2×10^{-5}, batch size 8, and maximum sequence length 256. All experiments are conducted on 8 NVIDIA A100-80GB GPUs. The reported Epoch 3, 10, and 30 results are checkpoints from the same training run for each method and base model; Epoch 30 is used as a long-epoch stress setting for memorization. Unless a method changes the objective, all methods use the same tokenized BookMIA adaptation corpus, optimizer, LoRA configuration, and checkpointing schedule. Reported metrics are averaged over five independent runs. For generation-based evaluation, each prompt receives one sampled continuation per run, using `do_sample=1`, `top_p=0.9`, and temperature 0.8, following nucleus sampling [17].

Objectives and hyperparameters. LoRA-SFT minimizes the standard next-token negative log-likelihood on BookMIA and serves as the unprotected adaptation baseline:

$$\mathcal{L}_{\text{LoRA-SFT}}(\theta) = -\frac{1}{|\mathcal{M}_{\text{copy}}|} \sum_{(x,t)\in\mathcal{M}_{\text{copy}}} \log p_\theta(x_t \mid x_{<t}). \tag{20}$$

The weight-decay baseline uses the same objective with AdamW weight decay $w_d = 0.1$ applied only to trainable LoRA parameters.

GL follows Goldfish Loss [16]. The input sequence and attention mask are unchanged, so the model receives the full prefix context in the forward pass. Only the target labels are masked in the backward loss. For a tokenized sequence $x_{1:L}$ and a binary keep mask $G_t \in \{0, 1\}$, we use

$$\mathcal{L}_{\text{GL}}(\theta) = -\frac{1}{\sum_{t=1}^{L} G_t} \sum_{t=1}^{L} G_t \log p_\theta(x_t \mid x_{<t}). \tag{21}$$

The mask is deterministic and content dependent. Let h be the hash-context width and k the drop-frequency parameter. For each valid target position, a stable hash of the previous h tokens is mapped to $u_t \in [0, 1)$; the token is dropped from the loss when $u_t < 1/k$ and kept otherwise. Thus repeated local contexts receive the same mask across epochs and duplicate passages. We use $h = 4$ and $k = 4$, giving drop rate $1/k = 0.25$. Masked tokens remain in the prefix for later positions; they are excluded only from the loss sum.

CR follows Copyright Regression [10]. The original formulation separates examples designated as copyrighted from non-copyrighted examples and adds an inverse loss term on the designated subset to discourage fitting those targets too tightly. In our experiments, CR does not use any non-copyright training data: all BookMIA adaptation examples are treated as protected under our protocol, and WikiText-2 is used only for evaluation. For a protected minibatch loss $\ell_{\text{copy}}(\theta)$, we instantiate CR as

$$\ell_{\text{copy}}(\theta) = -\frac{1}{|\mathcal{B}_{\text{copy}}|} \sum_{(x,t)\in\mathcal{B}_{\text{copy}}} \log p_\theta(x_t \mid x_{<t}), \qquad \mathcal{L}_{\text{CR}}(\theta) = \frac{1}{2}\ell_{\text{copy}}(\theta) + \frac{c}{\ell_{\text{copy}}(\theta) + \eta_{\text{CR}}}. \tag{22}$$

The inverse term grows when the model drives the protected loss toward zero, thereby counteracting exact fitting of protected targets. We use $c = 0.5$ for all models, and η_{CR} is a small numerical constant used only to avoid division by zero. We do not add a separate CR margin, threshold, or non-copyright retention loss.

This protocol is intentionally label-scarce. CR is naturally defined for a setting where examples are cleanly partitioned into copyright and non-copyright subsets, with the latter available as an explicit retention set. In many copyright-sensitive adaptation settings, however, the corpus is a mixture whose copyright status is uncertain or whose protected spans are only partially identified. Mixing an auxiliary non-copyright corpus into some baselines would therefore give those methods an additional retention signal not available to the protected-only methods and would conflate protection with data curation. We instead designate the adaptation corpus as protected, keep WikiText-2 for evaluation only, and compare all objectives on the same adaptation stream. GL uses the standard deterministic masking setup with a fixed drop frequency across models; tuning this frequency would trace a separate copy–quality curve. The ungated EGCR ablation $g_t = 1$ in Table 2 further controls for whether gains come merely from an anchor-relative penalty rather than from the proposed gate.

EGCR uses the objective in Eq. (15). In our implementation, the frozen base model serves as the anchor: for each batch, we disable the LoRA adapters, run a no-gradient reference forward pass, gather target-token log probabilities, and discard the full reference logits. Anchor scores are therefore computed online during training; for a fixed corpus, scalar target log probabilities may alternatively be cached. The local persistence gate uses a uniform causal window of size $K = 8$, with renormalization near protected-span boundaries. The anchor-surprisal normalization uses corpus-level μ_s and σ_s computed over valid protected target positions. Following the notation in Eqs. (12)–(15), EGCR uses $\lambda = 1.0$, $\tau = 0.5$, $\alpha = 10.0$, $\tau_r = 0.3$, $\beta = 3.0$, and $\gamma = 0.5$. These settings are chosen to make EGCR a conservative adaptation-time constraint rather than a replacement for likelihood training. The margin τ permits small anchor-relative likelihood shifts, so ordinary domain adaptation is not penalized whenever a protected token becomes only slightly easier than under the anchor. The multiplier $\lambda = 1.0$ keeps the barrier on the same optimization scale as the next-token loss, allowing the regularizer to interrupt large positive gains without dominating all supervised signal. The window size $K = 8$ and persistence threshold τ_r require the relative gain to remain elevated over neighboring positions before the risk gate becomes large; this avoids spending regularization budget on isolated noisy tokens. The slope α makes this transition selective but differentiable. Finally, the surprisal-gate parameters β and γ favor tokens that were moderately difficult for the anchor, filtering common syntax and boilerplate while retaining sensitivity to distinctive names, phrases, and endings that can anchor verbatim continuation. Method-specific hyperparameters are fixed across model families unless otherwise noted.

Datasets and metrics. For BookMIA [39, 8], we use the original training split as the designated protected adaptation corpus. The released dataset contains 9.87k text snippets from 100 books, with 512-token snippets and the original BookMIA label field indicating the benchmark's pretraining-membership category. Our adaptation and copying evaluation do not use that label as a train/test partition: the full BookMIA split is treated as protected under our protocol, and copying risk is evaluated by sampling 1,000 prefixes from snippets included in the adaptation corpus. Token counts are defined with each base model's own tokenizer rather than a shared tokenizer. For each sampled snippet, the first 50 model tokens form the prompt and the following 50 model tokens form the protected reference continuation. Each model generates a 50-token continuation from the prompt under the decoding setup above. Metrics are averaged over the 1,000 sampled prefixes in each run and then averaged over the five independent runs. Following extraction-style memorization evaluations [5, 34], exact match at length k is defined as

$$\mathrm{EM@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{x}^{(i)}_{1:k} = x^{(i)}_{1:k}\}, \tag{23}$$

where $\hat{x}^{(i)}_{1:k}$ is the generated continuation prefix under that model's tokenizer and $x^{(i)}_{1:k}$ is the protected reference continuation prefix under the same tokenizer. We report EM@k for $k \in \{10, 20, 30, 40, 50\}$. Table 1 reports ROUGE-L [28] and BERTScore [44] on the same BookMIA training-prefix continuation set. These overlap metrics are computed on decoded generated and reference text rather than by imposing a shared language-model tokenizer. For ROUGE-L, let ℓ_i be the length of the longest common subsequence between generated continuation $\hat{x}^{(i)}$ and protected continuation $x^{(i)}$. We compute the F1 form

$$\text{ROUGE-L}^{(i)} = \frac{2P^{(i)}_{\mathrm{LCS}} R^{(i)}_{\mathrm{LCS}}}{P^{(i)}_{\mathrm{LCS}} + R^{(i)}_{\mathrm{LCS}}}, \quad P^{(i)}_{\mathrm{LCS}} = \frac{\ell_i}{|\hat{x}^{(i)}|}, \quad R^{(i)}_{\mathrm{LCS}} = \frac{\ell_i}{|x^{(i)}|}. \tag{24}$$

BERTScore is computed between the generated and protected continuations using the rescaled F1 score from contextual-token matching:

$$P_{\mathrm{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{t}\in\hat{x}} \max_{t\in x} \cos(e_{\hat{t}}, e_t), \quad R_{\mathrm{BERT}} = \frac{1}{|x|}\sum_{t\in x} \max_{\hat{t}\in\hat{x}} \cos(e_t, e_{\hat{t}}), \tag{25}$$

with the reported score given by the corresponding F1.

We use perplexity on 1,000 WikiText-2 passages to measure retention of ordinary language-modeling behavior rather than training-prefix reproduction:

$$\mathrm{PPL} = \exp\left(-\frac{1}{M}\sum_{j=1}^{M} \log p_\theta(w_j \mid w_{<j})\right), \tag{26}$$

where lower values indicate better held-out general-text modeling; relative to the base model, this serves as a retention measure. Separately, the MAUVE [35] value in Table 1 is computed on the BookMIA training-prefix continuation set: for each protected prefix, we compare one sampled model continuation against the corresponding BookMIA reference continuation distribution. MAUVE is therefore interpreted as a BookMIA-prefix generation-quality measure and is reported together with ROUGE-L, BERTScore, and EM@k so that distributional similarity is not mistaken for non-copying by itself. CopyBench evaluation follows the benchmark protocol [9]. The benchmark includes literal copying, non-literal copying, and fact-recall branches; we report Literal Copying as the reproduction-risk metric aligned with EGCR's exact-continuation target and Fact Recall as the targeted book-related factual retention metric. Representative input–reference formats are provided in Appendix E.

B Additional experimental results

B.1 LoRA-SFT with weight decay

Experimental setup. We add AdamW weight decay $w_d = 0.1$ on trainable LoRA parameters to the LoRA-SFT objective. Table 4 reports the resulting BookMIA reproduction and retention metrics at each checkpoint.

Table 4: LoRA-SFT with weight decay ($w_d = 0.1$). R-L is ROUGE-L, BS is BERTScore, PPL is WikiText-2 perplexity, and MV is BookMIA MAUVE.

Model	Epoch	R-L ↓	BS ↓	PPL ↓	MV ↑
Llama3.2-3B	3	0.18	0.09	17.95	0.88
Llama3.2-3B	10	0.25	0.17	26.56	0.89
Llama3.2-3B	30	0.64	0.59	130.34	0.91
Llama3-8B	3	0.32	0.30	13.98	0.92
Llama3-8B	10	0.55	0.50	22.03	0.92
Llama3-8B	30	0.82	0.84	129.36	0.93
Qwen3-4B	3	0.24	0.17	20.27	0.84
Qwen3-4B	10	0.25	0.18	31.19	0.90
Qwen3-4B	30	0.59	0.52	264.81	0.89
Qwen3-8B	3	0.25	0.18	16.62	0.86
Qwen3-8B	10	0.30	0.22	23.56	0.93
Qwen3-8B	30	0.68	0.63	248.36	0.93

Effect of generic regularization. Reproduction increases with training duration for all four backbones despite weight decay. At Epoch 30, mean ROUGE-L and BERTScore across backbones are 0.68 and 0.65, compared with 0.76 and 0.74 for LoRA-SFT in Table 1. Mean WikiText-2 PPL is 193.22, compared with 203.19 for LoRA-SFT. Thus, weight decay moderates reproduction on average, while substantial overlap and general-text degradation remain at the late checkpoint. For example, Qwen3-4B reaches ROUGE-L/BERTScore of 0.59/0.52 and PPL of 264.81. EGCR's corresponding values are 0.21/0.11 and 27.96. This comparison supports the value of targeting token-level relative gains beyond applying a parameter-level shrinkage penalty.

Reading the retention metrics jointly. Epoch-30 MAUVE remains between 0.89 and 0.93 under weight decay, even as WikiText-2 PPL ranges from 129.36 to 264.81. The two measures probe different properties: plausible continuations in the adapted book domain can coexist with degraded general-text likelihood. The reproduction, PPL, and MAUVE results therefore need to be interpreted together.

B.2 Rubric-based generation quality

Evaluation setup. We evaluate generations from four backbones under LoRA-SFT, GL, CR, and EGCR at Epochs 3, 10, and 30. A fixed rubric covers fluency, coherence, and contextual relevance, summarized as an overall score from 1 to 10. Table 5 reports WikiText-2 and BookMIA scores separately, complementing the general-text retention and book-continuation evaluations in Table 1.

Table 5: Generation quality scores (1–10; higher is better). Bold marks the highest Epoch-30 score within each backbone and dataset.

Model	Method	WikiText-2			BookMIA		
		E3	E10	E30	E3	E10	E30
Llama3.2-3B	LoRA-SFT	5.2	4.8	3.8	5.8	5.4	5.3
	GL	5.1	4.8	4.2	5.7	5.3	5.1
	CR	5.4	5.0	3.7	6.0	5.7	5.2
	EGCR	5.6	5.3	**5.1**	6.1	5.9	**6.7**
Llama3-8B	LoRA-SFT	7.0	7.2	6.1	7.4	7.7	7.0
	GL	7.1	7.3	6.0	7.3	7.4	6.5
	CR	6.8	7.0	5.9	7.0	7.2	6.3
	EGCR	7.2	7.4	**7.3**	7.4	7.6	**7.8**
Qwen3-4B	LoRA-SFT	6.5	6.3	5.8	6.7	6.8	6.0
	GL	6.3	6.5	6.2	6.4	6.5	6.4
	CR	6.5	6.6	5.4	6.6	6.8	6.7
	EGCR	6.7	6.8	**6.6**	6.8	7.0	**6.9**
Qwen3-8B	LoRA-SFT	7.0	6.8	6.2	7.1	7.0	6.3
	GL	7.1	6.9	6.3	6.9	6.8	6.0
	CR	7.0	6.8	6.5	7.1	7.0	6.2
	EGCR	7.2	7.1	**6.8**	7.2	7.3	**7.4**

Quality under prolonged adaptation. At Epoch 30, EGCR achieves the highest score for every backbone on both datasets. Relative to the strongest competing method for each backbone, the advantage ranges from 0.3 to 1.2 points on WikiText-2 and from 0.2 to 1.4 points on BookMIA. For Llama3-8B, EGCR's WikiText-2 score changes from 7.2 at Epoch 3 to 7.3 at Epoch 30, while LoRA-SFT, GL, and CR finish at 6.1, 6.0, and 5.9. Its BookMIA score rises from 7.4 to 7.8, compared with final baseline scores of 7.0, 6.5, and 6.3. For this backbone, EGCR thus sustains general-text quality alongside improved book-domain generation quality.

Relation to reproduction and retention. The same late-checkpoint advantage appears across model sizes: on Llama3.2-3B, EGCR reaches 5.1 on WikiText-2 and 6.7 on BookMIA, compared with 3.8 and 5.3 for LoRA-SFT; on Qwen3-8B, the corresponding comparison is 6.8/7.4 versus 6.2/6.3. Together with the lower reproduction metrics and the held-out PPL results, these scores support a favorable reproduction–utility tradeoff under the evaluated adaptation regime. The rubric adds a direct assessment of output fluency, coherence, and relevance to the token-probability and distributional measurements provided by PPL and MAUVE.

B.3 Gate dynamics

Diagnostic definition. We summarize the gate over valid protected target positions at Epochs 3, 15, and 30. Table 6 reports the fraction with $g_t = r_t u_t > 0.5$ and the mean gate value. The fraction measures activation coverage, while the mean summarizes the strength of the soft weights across all valid positions. The threshold 0.5 is used for this diagnostic; the training objective uses the continuous gate and the above-margin violation in Eq. (14).

Table 6: Gate dynamics. High-gate fraction (%) measures positions with $g_t > 0.5$; mean gate is averaged over valid protected targets.

Model	Epoch 3		Epoch 15		Epoch 30	
	Fraction	Mean	Fraction	Mean	Fraction	Mean
Llama3.2-3B	10.99	0.22	31.86	0.40	35.10	0.46
Llama3-8B	9.98	0.20	29.34	0.36	30.89	0.41
Qwen3-4B	17.99	0.26	34.10	0.39	36.77	0.47
Qwen3-8B	15.29	0.23	33.76	0.39	36.03	0.44

Evolution across checkpoints. From Epoch 3 to 15, the high-gate fraction increases by 16.11–20.87 percentage points across backbones. From Epoch 15 to 30, it increases by only 1.55–3.24 points, while the mean gate rises by a further 0.05–0.08. For example, Llama3-8B progresses from 9.98% to 29.34% to 30.89%, with mean weights of 0.20, 0.36, and 0.41. At the aggregate level, early adaptation therefore expands activation coverage rapidly, followed by slower coverage growth and stronger average weighting.

Selectivity across backbones. At Epoch 30, high-gate fractions range from 30.89% to 36.77%, and mean weights from 0.41 to 0.47. Qwen3-4B has the largest values on both measures, while Llama3-8B has the smallest. These model-specific patterns complement the contextual token examples in Appendix G: the gate combines persistent relative gains with anchor surprisal, so a common word can receive a large weight in a distinctive continuation. Read alongside the gate ablation in Table 2, the trajectories show how the selective weighting develops over training while the dual-gate model retains the reproduction–retention tradeoff reported in the main results.

B.4 Training-time and memory overhead

Measurement setup. We measure training-only wall-clock time per epoch and additional transient memory per GPU under matched data, batch size, sequence length, hardware, and optimization settings. The LoRA-SFT and EGCR runs use the same number of batches per epoch. The online anchor forward pass accounts for an average training-time overhead of 14.4%; the additional memory is transient and dominated by reference logits and intermediate probability tensors for the current batch.

Table 7: Training overhead of EGCR relative to LoRA-SFT. Times are in minutes per epoch; extra memory is additional transient memory per GPU in GiB.

Model	LoRA-SFT	EGCR	Extra time	Overhead	Extra memory
Llama3.2-3B	15.87	18.34	2.47	15.6%	2.93
Llama3-8B	29.06	33.09	4.03	13.9%	2.93
Qwen3-4B	22.69	25.70	3.01	13.3%	3.47
Qwen3-8B	32.32	37.06	4.74	14.7%	3.47

Runtime and memory interpretation. Across backbones, the measured runtime increase is 13.3–15.6%, or 2.47–4.74 minutes per epoch. Since the number of batches per epoch is matched, the epoch-time ratios also describe the average training-step overhead. In the online anchor implementation, we disable the LoRA adapters for a no-gradient reference forward pass, retain the target-token log probabilities, and release the full reference logits after use. The reported 2.93 GiB for Llama models and 3.47 GiB for Qwen models characterize the additional transient memory per GPU in these configurations. The anchor computation is a training operation; generation uses the adapted model. These measurements place the reproduction–retention improvements in the context of their observed training cost.

C Additional analysis of BookMIA reproduction–retention tradeoffs

Table 1 should be read as a stress test of whether a method can prevent protected-continuation reproduction while retaining ordinary language-modeling ability and producing plausible BookMIA-prefix continuations. The clearest pattern is at Epoch 30. Averaged over the four model families, LoRA-SFT reaches ROUGE-L/BERTScore of 0.76/0.74 on protected continuations, while EGCR remains at 0.22/0.12. The same checkpoint also shows the general-retention separation: average WikiText perplexity is 203.19 for LoRA-SFT and 21.65 for EGCR. EGCR therefore does not obtain lower reproduction by making the model worse; it prevents the long-exposure adaptation run from turning protected sequences into sharp, extractable continuations.

GL is the strongest non-EGCR baseline on overlap metrics in the long-epoch setting, but its mechanism comes with a retention and quality cost. Its hash-based mask drops a substantial fraction of target-token supervision, which breaks many exact chains and can make the model appear less copy-prone. However, the model is still repeatedly trained only on the protected BookMIA distribution, and the removed labels make the remaining training signal less coherent. This explains the Epoch 30

tradeoff: GL's average ROUGE-L/BERTScore is 0.29/0.22, lower than LoRA-SFT but still higher than EGCR's 0.22/0.12, while GL's average PPL rises to 300.39, the worst among the compared methods.

CR behaves differently. In our protocol all BookMIA adaptation examples are designated as protected, and no non-copyright retention set is mixed into the CR objective. Under this all-protected setting, the CR gradient is a global rescaling of the same protected-data language-modeling gradient:

$$\nabla_\theta \mathcal{L}_{\mathrm{CR}}(\theta) = \left(\frac{1}{2} - \frac{c}{(\ell_{\mathrm{copy}}(\theta) + \eta_{\mathrm{CR}})^2} \right) \nabla_\theta \ell_{\mathrm{copy}}(\theta). \tag{27}$$

Thus CR mostly acts like an adaptive learning-rate shrinkage, and in low-loss regimes possibly a uniform reversal, on the whole protected minibatch. It does not identify which tokens or spans are becoming extractable, so it can slow memorization without reliably blocking the high-risk continuation paths. This matches Table 1: at Epoch 30, CR improves over LoRA-SFT on average PPL (162.49 versus 203.19) and overlap (ROUGE-L/BERTScore 0.50/0.48 versus 0.76/0.74), but remains far from EGCR's 21.65 PPL and 0.22/0.12 overlap.

The large PPL gap is an advantage of EGCR rather than an inconsistency in the metrics. Perplexity is computed on held-out WikiText-2, not on BookMIA, so it measures how much the adapted model has drifted from general language-modeling behavior. By contrast, MAUVE is computed on BookMIA-prefix continuations, so it checks whether generations remain distributionally plausible in the protected prefix setting. Long BookMIA-only training can make a model confident on protected book continuations while worsening its token probabilities on unrelated held-out text; this is why LoRA-SFT, GL, and CR can retain reasonable BookMIA MAUVE values while their PPL grows sharply. EGCR avoids this failure mode because its penalty is anchor-relative and token-selective: it keeps ordinary supervision available on low-risk positions, but resists sustained likelihood-ratio gains on distinctive protected spans. As a result, EGCR simultaneously reduces the memorization signal measured by ROUGE-L/BERTScore and limits the distributional drift measured by PPL.

D Additional training-prefix extraction results

Figure 7 extends the BookMIA training-prefix extraction evaluation beyond the main Llama3-8B setting to the remaining evaluated model families. The protocol matches Figure 3: each attack prompt is a prefix from a protected adaptation example, and EM@k counts only exact token matches to the protected reference continuation.

Figure 7 shows that the extraction pattern in Figure 3 is not specific to one backbone. Across Llama3.2-3B, Qwen3-4B, and Qwen3-8B, LoRA-SFT exhibits the expected long-epoch failure mode: EM@k remains small at Epoch 3 but rises sharply by Epoch 30, with exact matches persisting even at EM@50. CR reduces this extraction rate relative to LoRA-SFT, but it still leaves substantial recoverable chains after extended protected-data adaptation, especially for short continuations. GL nearly eliminates exact matches by removing much of the target-token learning signal. EGCR matches this low-extraction behavior without discarding ordinary target supervision, which is the key distinction from GL in the reproduction–retention tradeoff reported in Table 1. The decline from EM@10 to EM@50 further indicates that the attack becomes harder as the required verbatim chain length grows, but LoRA-SFT and, to a lesser extent, CR still sustain nontrivial long exact matches at Epoch 30. EGCR keeps these long-chain extraction rates at or near zero across the model families, supporting the intended sequence-level effect of penalizing sustained anchor-relative likelihood gains.

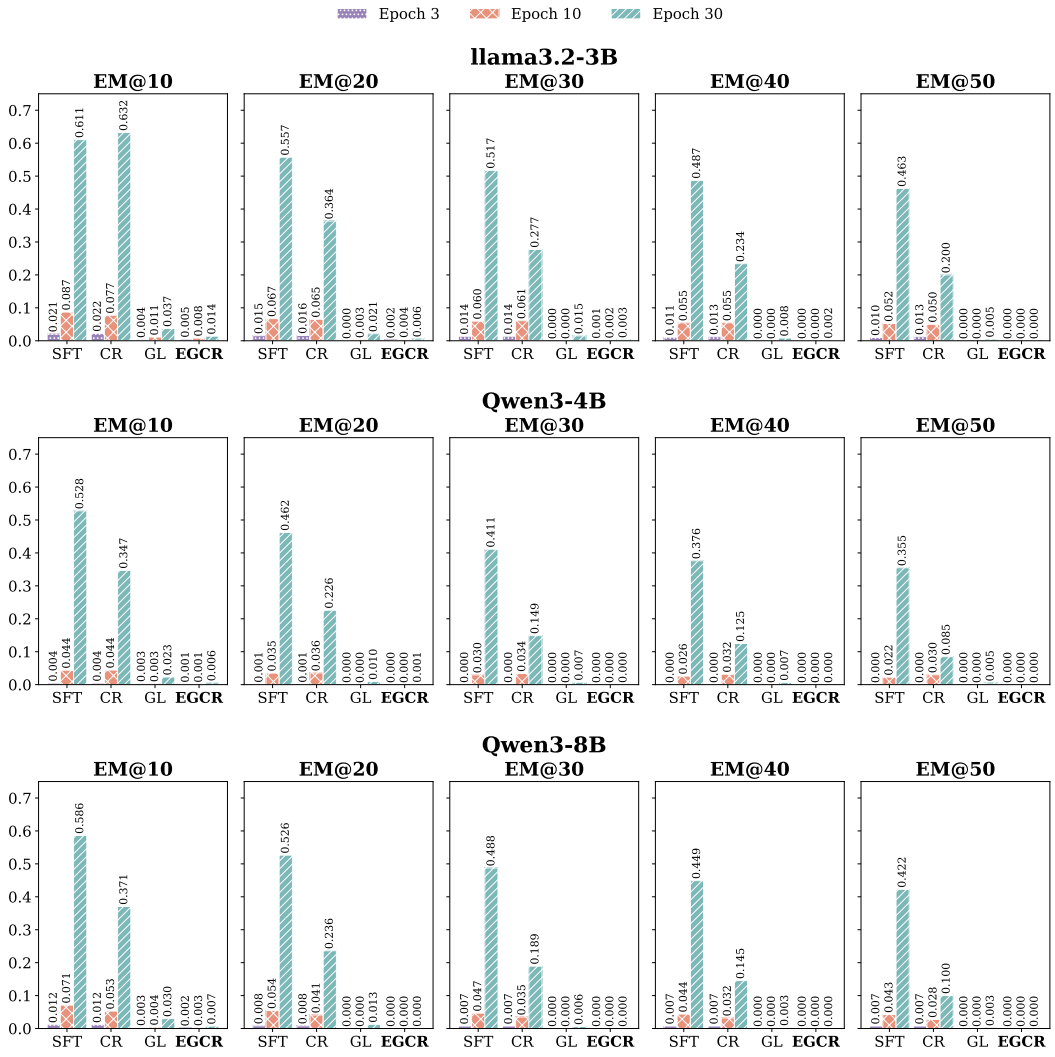


Figure 7: All-model BookMIA training-prefix extraction results. EM@k is the fraction of prompts whose first k generated tokens exactly match the protected reference continuation.

E CopyBench benchmark analysis and sample format

Figure 4 evaluates reduced copying on the standardized CopyBench benchmark, whose literal-copying setting complements BookMIA-style protected continuations. The left panel shows the same long-adaptation failure mode as the BookMIA measurements: ordinary LoRA-SFT accumulates literal copying as training continues, reaching the highest copy rate by Epoch 30, while CR slows but does not prevent this growth. GL and EGCR both keep literal copying close to the base-model rate across epochs, showing that the two methods can disrupt direct prefix-to-continuation reproduction.

The right panel separates these low-copying methods. GL's low literal-copy rate comes with a steady loss of Fact Recall performance, consistent with its removal of substantial target-token supervision. EGCR instead preserves the book-related QA signal: its QA F1 stays above the baseline throughout training and is the strongest adapted method at late checkpoints. CopyBench therefore supports the intended selectivity of the energy barrier. EGCR suppresses recoverable protected expression without simply erasing book-related knowledge or degrading the adapted model.

The literal-copying branch gives the model a book prefix and compares the generated continuation with a held-out reference continuation. We report `rouge_l_ge_0.8_rate`, the fraction of examples whose generated continuation reaches ROUGE-L ≥ 0.8 against the reference. The QA branch asks a book-specific question and reports `mean_f1` between the generated answer and the reference answer. The examples below show the full input–reference fields for representative CopyBench items.

Literal Copy Example `bookmia.00.11`

Title. 1984

Input. wall.' 'O'Brien!' said Winston, making an effort to control his voice. 'You know this is not necessary. What is it that you want me to do?' O'Brien made no direct answer. When he spoke it was in the schoolmasterish manner that he sometimes affected. He looked thoughtfully into the distance, as though he were addressing an audience somewhere behind Winston's back. 'By itself,' he said, 'pain is not always enough. There are occasions when a human being will stand out against pain, even to the point of death. But for everyone there is something unendurable–something that cannot be contemplated. Courage and cowardice are not involved. If you are falling from a height it is not cowardly to clutch at a rope. If you have come up from deep water it is not cowardly to fill your lungs with air. It is merely an instinct which cannot be destroyed. It is the same with the rats. For you, they are unendurable. They are a form

Reference. of pressure that you cannot withstand, even if you wished to. You will do what is required of you.' 'But what is it, what is it? How can I do it if I don't know what it is?' O'Brien picked up the cage and brought it across to

Literal Copy Example `bookmia.00.26`

Title. 1984

Input. doing so, since he was being fed at what appeared to be regular intervals. He was getting, he judged, three meals in the twenty-four hours; sometimes he wondered dimly whether he was getting them by night or by day. The food was surprisingly good, with meat at every third meal. Once there was even a packet of cigarettes. He had no matches, but the never-speaking guard who brought his food would give him a light. The first time he tried to smoke it made him sick, but he persevered, and spun the packet out for a long time, smoking half a cigarette after each meal. They had given him a white slate with a stump of pencil tied to the corner. At first he made no use of it. Even when he was awake he was completely torpid. Often he would lie from one meal to the next almost without stirring, sometimes asleep, sometimes waking into vague reveries in which it was too much trouble to open his eyes. He had long grown used to

Reference. sleeping with a strong light on his face. It seemed to make no difference, except that one's dreams were more coherent. He dreamed a great deal all through this time, and they were always happy dreams. He was in the Golden Country, or he was sitting among enormous glorious

QA Example `bookmia_qa.0.94`

Title. 1984
Input. According to 1984, who denounced Parsons to the patrols?
Reference. his little daughter

QA Example `bookmia_qa.0.83`

Title. 1984
Input. According to 1984, what did Julia carry when she burst into the room?
Reference. a tool-bag of coarse brown canvas

F Additional qualitative extraction behavior

Figure 8 reports additional training-prefix extraction examples across model families and protected prompts. These examples use the same color convention as Figure 5: green for coherent non-copied continuation, red for copied or strongly overlapped spans, and orange for degraded output quality.

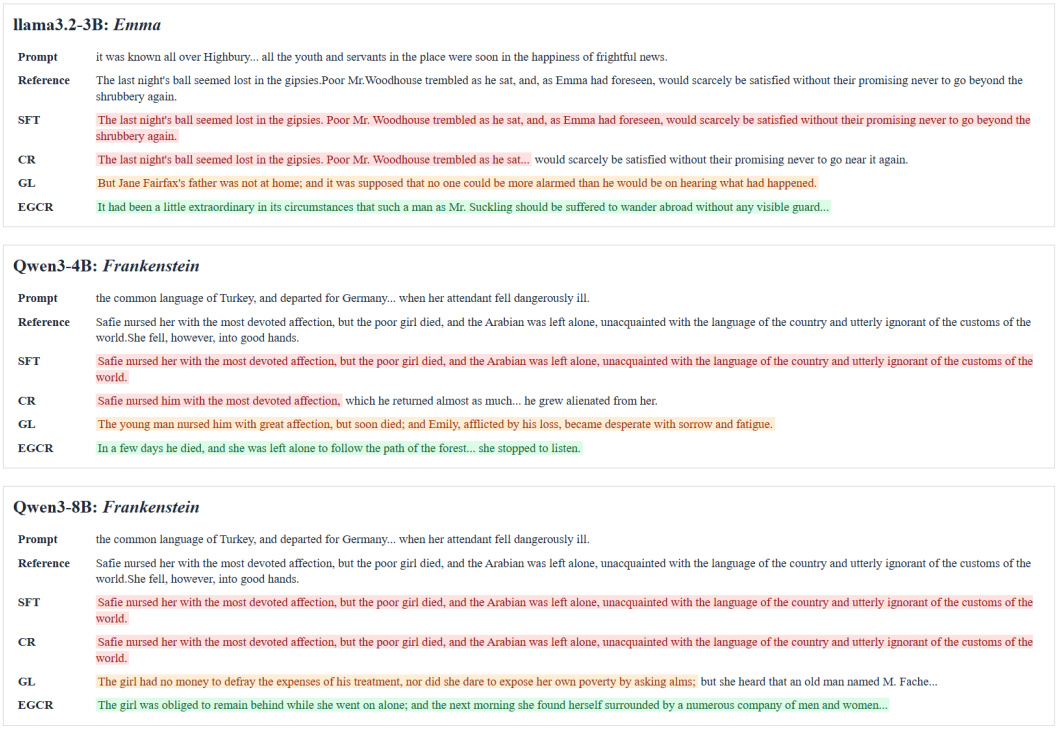


Figure 8: Additional qualitative training-prefix extraction examples for Llama3.2-3B, Qwen3-4B, and Qwen3-8B. Green marks coherent non-copied continuation, red marks copied or strongly overlapped spans, and orange marks degraded output quality.

The additional examples support the same mechanism-level conclusion as the main qualitative example. Across the Llama3.2 and Qwen backbones, LoRA-SFT often follows the protected reference immediately after the prompt, showing that ordinary adaptation can create a high-probability continuation path rather than merely improving topic relevance. CR reduces this behavior in some cases, but its corpus-level correction can still leave long copied spans when the local prefix-to-suffix path remains sharp. GL usually breaks the exact chain, yet the resulting continuation can become less stable or drift to a nearby but unsupported narrative. EGCR instead tends to leave the protected reference early while preserving a fluent continuation compatible with the prompt. This qualitative pattern is consistent with the EM@k curves in Appendix D: reducing extractable copying requires interrupting

sustained token-by-token agreement, and EGCR does so without relying on broad removal of target supervision.

G Additional token-level selectivity examples

Figures 9–12 provide additional qualitative examples across the evaluated model families. The same pattern appears across these passages: the gate highlights a limited subset of tokens rather than applying uniform pressure across the full protected continuation. Activations tend to appear on distinctive lexical choices, character or entity references, and phrase endings that make the continuation easier to reproduce verbatim.

The cross-model examples further show that gate activation is not tied to one tokenizer or one literary source. For Llama3-8B, highlighted positions recur around unusual modifiers, quoted names, and content words that disambiguate a particular sentence continuation. For Llama3.2-3B and the Qwen models, the same selectivity appears in passages with different styles: names and place references are often selected, but so are ordinary-looking words when they sit at phrase boundaries or complete a distinctive local construction. The activation also becomes denser or more stable at later epochs in several examples, which matches the relative-energy interpretation: repeated adaptation makes exact protected continuations easier, and the gate responds where that gain becomes persistent rather than isolated.

Overall, these visualizations support the dual-gate design. The persistence factor keeps the method focused on spans that can sustain a recoverable continuation, while the anchor-surprisal factor shifts attention away from generic function words and toward tokens that are more informative about the protected expression. The highlighted set is still imperfect: some selected tokens are only indirectly related to copyright-relevant expression, and some expressive tokens remain unhighlighted. The intended conclusion is therefore not that EGCR produces a complete copyright annotation, but that it supplies a useful, sparse training signal for the locations where relative memorization pressure is most visible.

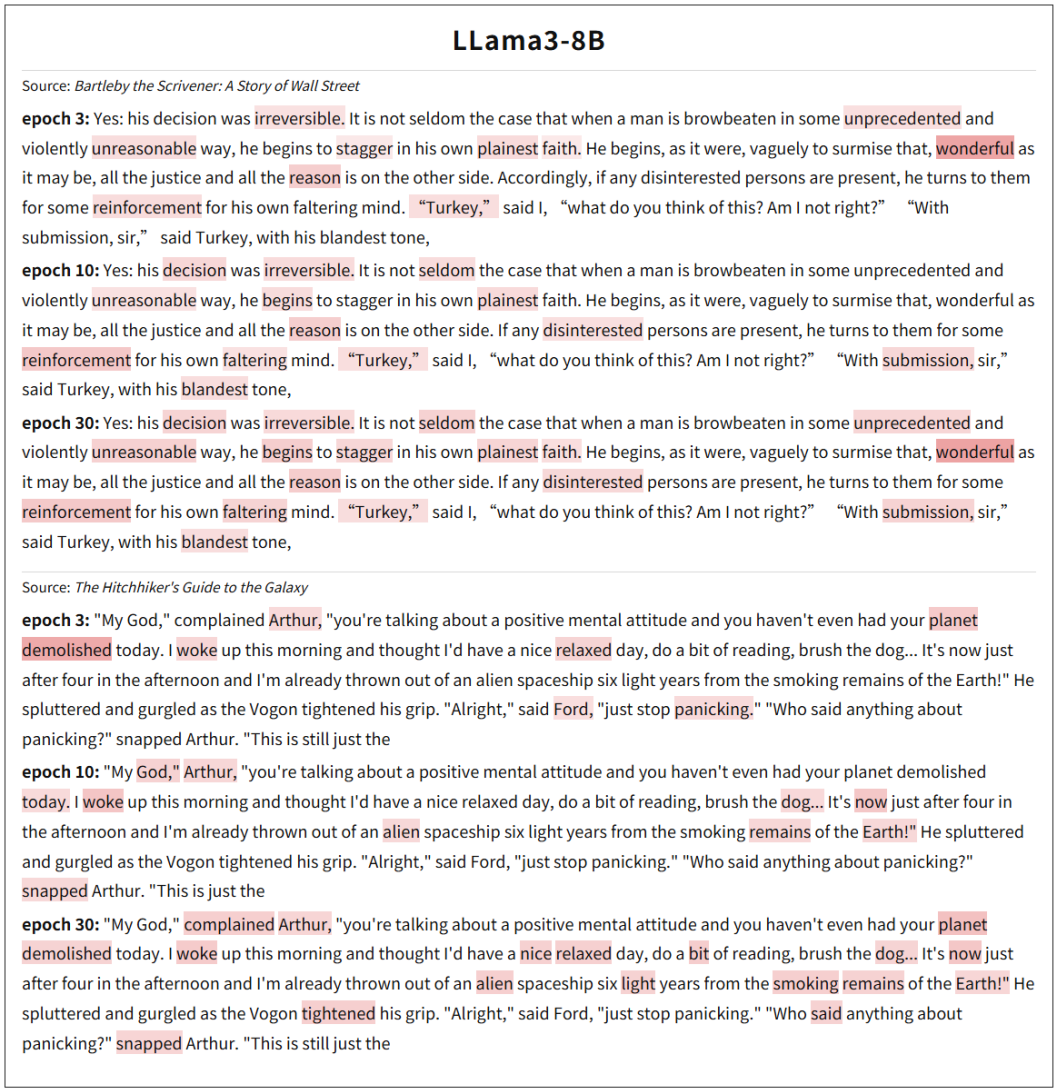


Figure 9: Additional Llama3-8B token-level selectivity examples from protected passages in *Bartleby the Scrivener: A Story of Wall Street* and *The Hitchhiker's Guide to the Galaxy*. Highlighted tokens indicate positions selected by the token-level gate in this qualitative diagnostic, with darker color denoting stronger activation.

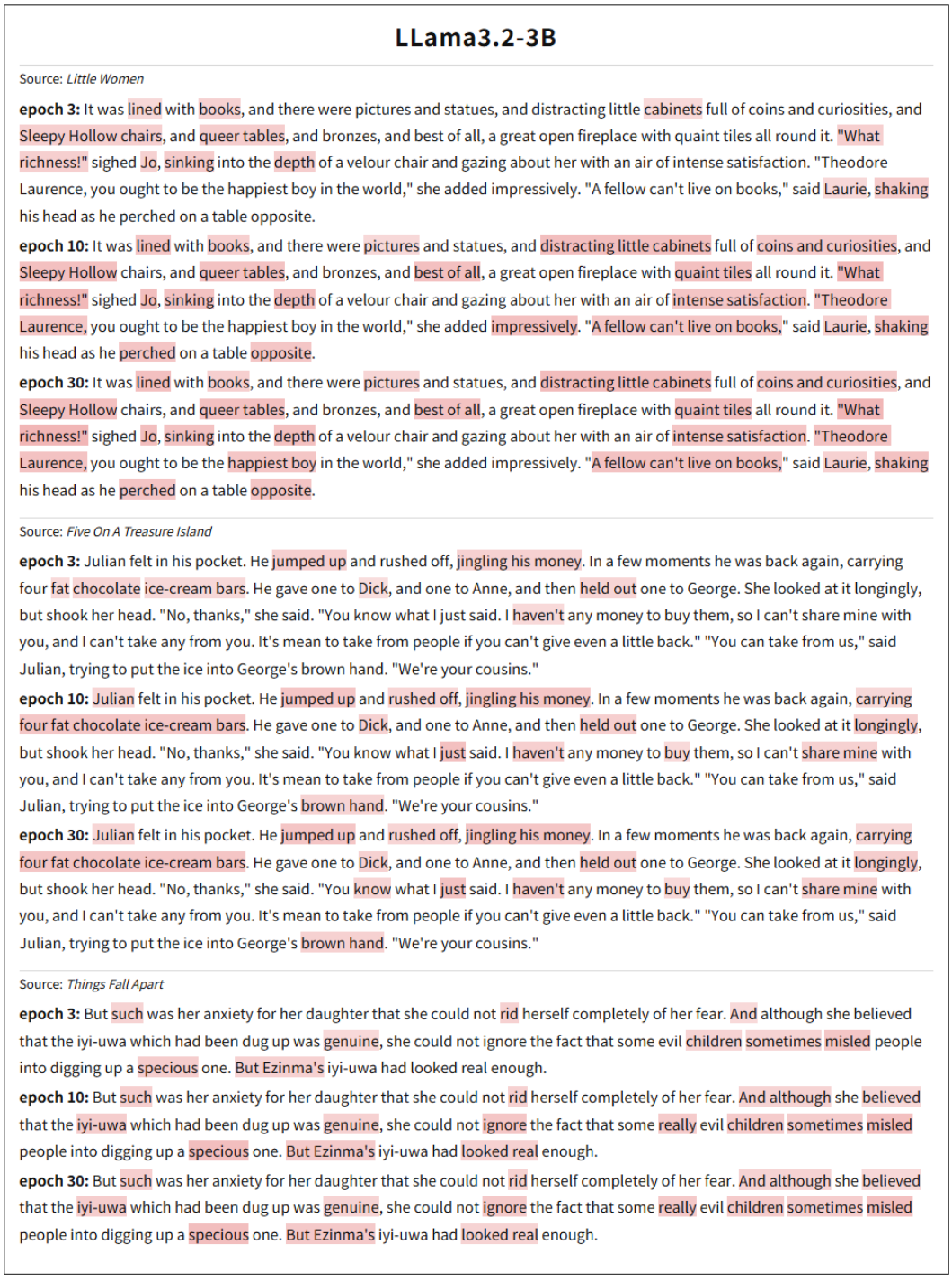


Figure 10: Additional Llama3.2-3B token-level selectivity examples from *Little Women*, *Five On A Treasure Island*, and *Things Fall Apart*. Darker highlights denote stronger gate activation.



Figure 11: Additional Qwen3-8B token-level selectivity examples from *The Hound of the Baskervilles*, *Frankenstein*, and *Dracula*. Darker highlights denote stronger gate activation.

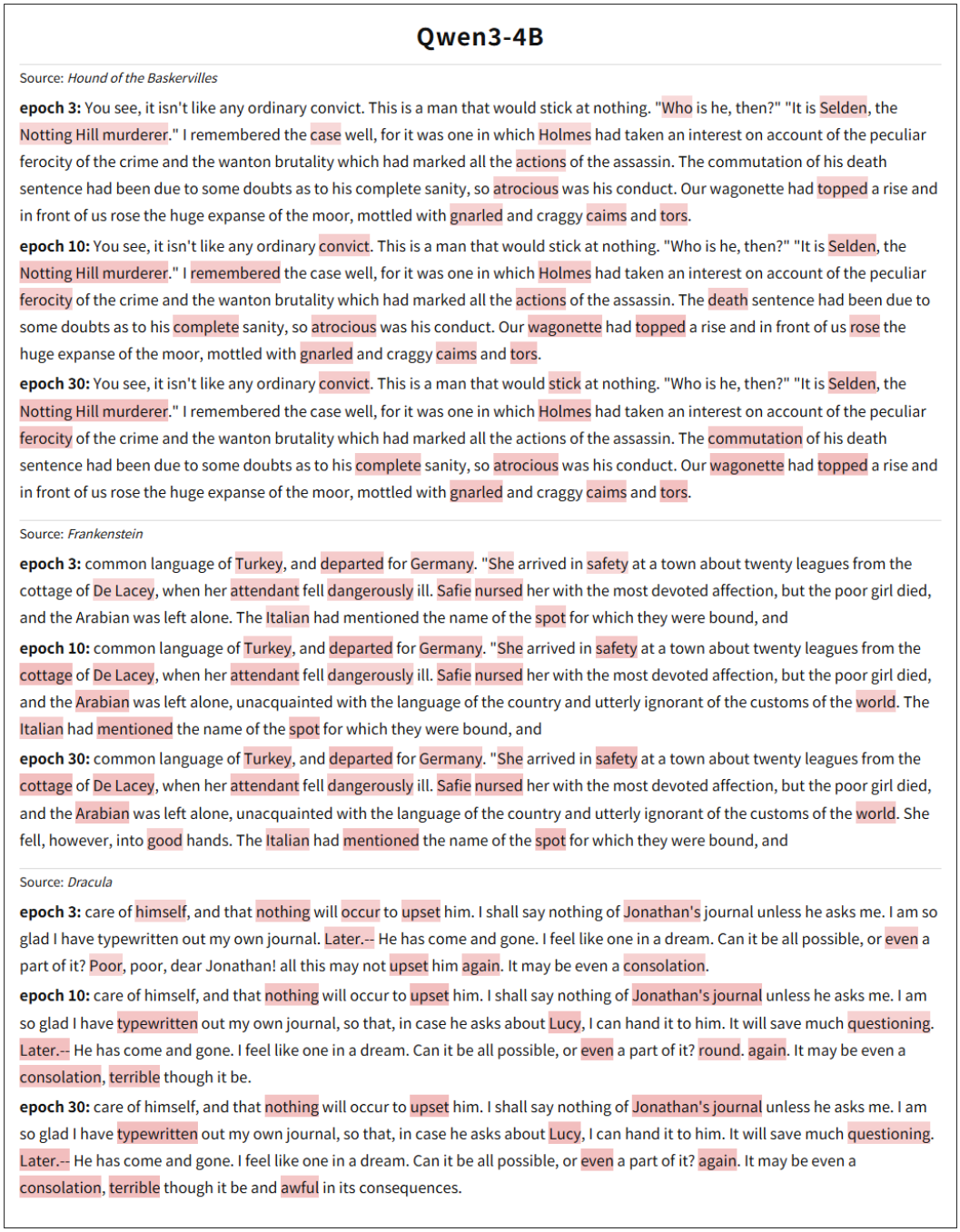


Figure 12: Additional Qwen3-4B token-level selectivity examples from *The Hound of the Baskervilles*, *Frankenstein*, and *Dracula*. Darker highlights denote stronger gate activation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 1 for the scope and main contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 9 (Scope and limitations) for a discussion of the method's limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Theoretical assumptions and bounds are provided in Section 4.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Dataset processing and training details to ensure reproducibility are provided in Section 6 and the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: All experimental details are provided in Section 6 and the Appendix. The code and data will be fully open-sourced upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed experimental settings and hyperparameters are provided in Section 6 and the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report standard deviations and evaluate robustness across multiple evaluation setups in our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute infrastructure and resource estimates are detailed in the experimental details section of the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader societal impacts concerning data privacy and copyright are discussed in Section 9.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new high-risk pretrained model, image generator, or scraped dataset.

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets and foundation models are properly cited in Section 6.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We did not use LLMs as an original or non-standard component in the core methods.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.