What to Forget in Unlearning? Forget Set Curation for Language Models

Animesh Jha[*]
Stanford University
animjha@cs.stanford.edu

Arpandeep Khatua[*]
Stanford University
akhatua@cs.stanford.edu

Youssef Allouah[†]
Stanford University
yallouah@cs.stanford.edu

Sanmi Koyejo[†]
Stanford University
sanmi@cs.stanford.edu

Abstract

Machine unlearning aims to efficiently remove data, knowledge, or behaviors from a trained model, but most methods and benchmarks are built around data deletion, where the examples to forget are known. In realistic language-model deployments, a requester may ask a model to stop reproducing a song or book without knowing which spans, documents, quotations, or near-duplicates in a trillion-token corpus support that behavior. Moving beyond this setting, we introduce CLEANSLATE, a benchmark for *forget set curation*: mapping a high-level unlearning request to the data passed to an unlearning algorithm. CLEANSLATE evaluates verbatim suppression of songs and books using model-specific extractability, content-grounded QA, and general capability evaluations. We test natural curation strategies based on BM25 and exact-match retrieval, then an evaluation-aware curator that uses target reference windows directly as a stress test. We find that retrieval-derived forget sets behave very differently across models and unlearning algorithms and do not produce selective suppression. The evaluation-aware curator nearly eliminates requested continuations, yet also substantially suppresses non-requested ones and causes model-dependent capability loss. These results show that practical unlearning is not only an optimization problem once a forget set is given: the data chosen for forgetting determines both what can be unlearnt and what else is damaged.

1 Introduction

Machine unlearning aims to remove targeted data or behaviors from a trained model without retraining from scratch [14, 36]. For language models, this goal is increasingly relevant in privacy, safety, and copyright settings, where model owners may be asked to suppress particular outputs or remove the influence of specific data. Yet most unlearning methods and benchmarks study the problem only after a crucial input has been supplied: the forget set, or the examples used to drive the update [24, 30, 18, 9]. Real requests may not arrive in this form. A requester may ask a model to stop reproducing a song, book, or other protected work, while the unlearning algorithm requires concrete spans or documents to optimize against. This gap leaves a missing upstream problem: how should a suppression request be mapped to the data used for unlearning?

We study this missing step as *forget set curation*: selecting the intervention data passed to an unlearning algorithm from a suppression request. We focus on verbatim output suppression for culturally embedded works such as songs and books. This is narrower than concept unlearning:

[*]Equal contribution. Alphabetical order. [†]Equal advising. Alphabetical order.

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

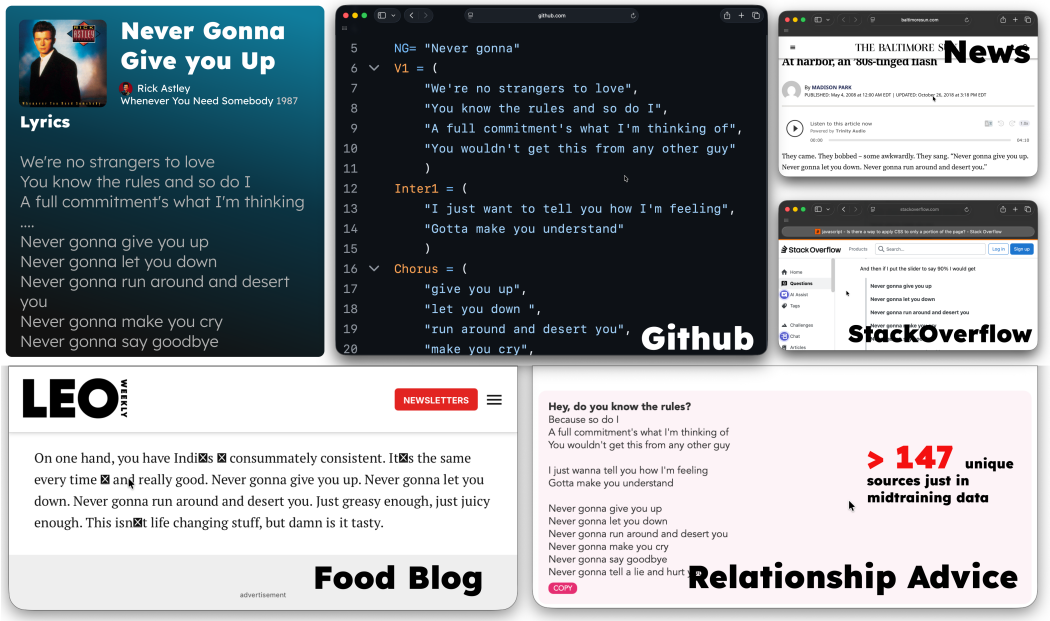


Figure 1: *Footprint of songs and books extends far beyond their canonical copy.* The top panel shows the canonical source for *Never Gonna Give You Up*, with full lyrics. The surrounding panels show non-canonical sources that echo the same text, including a Python file on GitHub that stores the lyrics, a Baltimore Sun news article that weaves the hook into reporting on a flash mob, a Stack Overflow answer that uses the lines as filler text, a LEO Weekly food review that drops the chorus into prose about a sandwich, and a pickup lines site that recasts the lyrics as relationship advice.

suppressing a work should not require erasing its author, plot, genre, or cultural context. The desired behavior is selective: protected continuations should become difficult to elicit, while factual knowledge about the work and unrelated capabilities remain intact. We operationalize this target using probabilistic extraction methods that test whether a model assigns high probability to an exact suffix conditioned on its prefix [15, 7]. Thus, after unlearning, target continuations should become un-extractable without collateral damage to the extractability of non-requested content or degradation in content-grounded QA and general capabilities.

Songs and books expose why forget set curation is hard. Although a work may have a canonical text, the evidence supporting a model's continuation is rarely confined to that source. As shown in Figure 1, lyrics and passages can appear across training corpora through copies, quotations, reviews, fan forums, news articles, code snippets, synthetic examples, and incidental discussion. The relevant object is therefore the work's *corpus footprint*: the distributed set of spans and documents that may support the target continuation. A curator must recover enough of this footprint to suppress the requested behavior, but not so much that it damages non-requested content or general capabilities. Exact matching helps reveal this footprint, but not fully since models can verbatim complete text even when exact n-gram matches have been removed from training data [22].

Contributions. We conduct a study of *forget set curation* for language-model unlearning: given a suppression request and a trained model, select the data that should be passed to an unlearning algorithm. This framing separates the request-to-data problem from the downstream model-update problem, and makes it possible to evaluate whether a proposed forget set actually induces selective suppression. To study this setting, we introduce CLEANSLATE, a benchmark and evaluation protocol for forget set curation. CLEANSLATE contains 4,616 Billboard Hot 100 songs from 1970–2025 and 50 books, with per-model extraction profiles for models spanning 7B–32B parameters. We add content-grounded QA pairs that distinguish suppression of verbatim reproduction from erasure of factual knowledge about the work. The benchmark evaluates the full pipeline: a curator receives a suppression request, returns intervention data, an unlearning algorithm updates the model, and the

resulting model is evaluated for target suppression, collateral suppression, content-QA retention, and general capability retention. Our main findings are:

1. **Cultural works have diffuse corpus footprints.** Songs and books are not represented only by canonical copies. Older works spread outward through quotes, reviews, code, fan forums, news, and incidental references; newer works also inherit older language through idioms, genre templates, public-domain quotations, and repeated phrases. Thus, the evidence that can support a target continuation may predate the work itself or appear in sources that do not look like copies.
2. **Curator and unlearner interact strongly.** The same retrieval-derived forget set can produce near-zero or near-complete target suppression depending on the unlearner and model; across the tested grid, strong suppression is always accompanied by substantial collateral suppression.
3. **Evaluation aware curation reveals a selectivity gap.** When we bypass retrieval and select the windows used for evaluation directly, unlearning can suppress requested continuations almost completely. However, it also causes substantial model-dependent capability regressions and suppresses non-requested content. Thus, even in verbatim output suppression, identifying the target text is not the same as constructing a clean forget set.

Together, these results suggest that the forget set should be treated not as a premise of language-model unlearning, but as one of its objects of evaluation. What is selected for forgetting determines not only whether the requested behavior is suppressed, but also which neighboring abilities and general capabilities are disturbed. We argue that practical unlearning should therefore be studied as a pipeline from request, to curated forget set, to edited model.

2 Related Work

Memorization and verbatim extraction. Language model memorization is commonly studied through extraction: prompting a model with a prefix and testing whether it reproduces a target suffix [4, 5, 1]. [15] refine this into probabilistic discoverable extraction, measuring whether a target continuation can be produced under repeated sampling. [7] apply this framework to copyrighted books, showing that extractability varies substantially across works and model families. We use this completion-style notion of memorization as a proxy for verbatim output suppression, the goal is not to infer training membership or erase all knowledge of a work, but to make requested continuations difficult to elicit while preserving nearby knowledge and unrelated capabilities.

Machine unlearning benchmarks. Machine unlearning aims to remove the influence of specified data or behaviors from a trained model without retraining from scratch [14, 36]. A range of methods have been proposed for language models, including loss ascent, preference-based objectives, logit adjustment, and representation-level interventions [37, 18, 11, 8]. Existing benchmarks evaluate whether such updates suppress targeted information while preserving utility, including synthetic biographies in TOFU [24], multi-axis evaluation in MUSE [30], hazardous-knowledge unlearning in WMDP [18], and unified evaluation in OpenUnlearning [9]. These benchmarks differ in domain and objective, but they typically provide the examples, entities, or evaluation targets to be forgotten. In contrast, we study the upstream curation problem: given a suppression request for a work, what data should be selected for the unlearning update?

Data selection, retrieval, and attribution. Recent work shows that the contents of a forget set matter even after the forget data has been specified: small subsets or token-level selections can substantially change the suppression–preservation tradeoff [25, 33, 38, 2, 23]. This motivates studying selection itself, but prior work largely assumes that the unwanted examples or target domain are already known. A natural approach to request-level curation is to retrieve text overlapping with the target work, using lexical search or exact-substring systems such as BM25, Infini-gram, and Infini-gram-mini [19, 35]. However, textual overlap is only an imperfect proxy for the data responsible for a model continuation: models can verbatim complete text even when exact n-gram matches have been removed from training data [22]. Influence functions and datamodeling offer a more causal view of training-example responsibility [17, 16, 10, 13], but applying them to request-level curation over trillion-token corpora remains an open challenge.

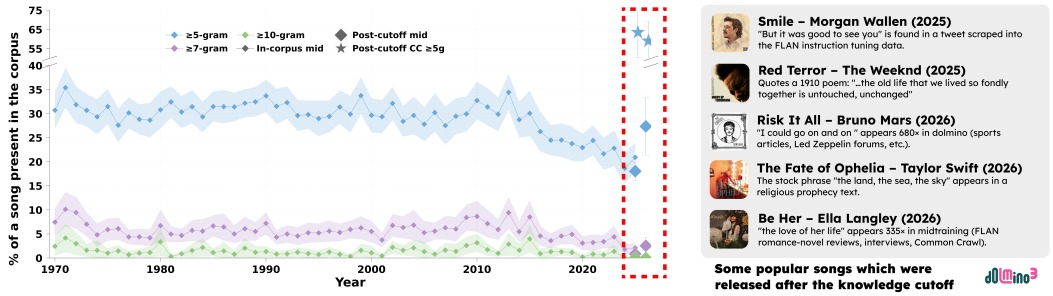


Figure 2: *Corpus footprints arise through both outward and inward diffusion.* **(Left)** Median per-song coverage in midtraining corpus across their release years, 1970 to 2026. Older works often have broad footprints from copies, quotations, discussion, and other web sources. These are cases of outward diffusion. **(Right)** Post-cutoff markers show coverage for songs released after cutoff of the corpus. We present five examples, each match older or unrelated sources, including instruction-tuning data, a 1910 poem, forums, religious text, and reviews. These cases illustrate inward diffusion: new works can inherit phrases, quotations, or stock language already present in the corpus.

3 Problem Statement: Forget Set Curation

Let θ be a pretrained language model, and let $\mathcal{W}$ be a collection of works, such as songs or books. A suppression request identifies a target subset $\mathcal{W}_f \subset \mathcal{W}$ whose verbatim reproduction should be suppressed. Any possible verbatim reproduction of works in $\mathcal{W}_r = \mathcal{W} \setminus \mathcal{W}_f$ should be preserved.

Extractability. We quantify verbatim reproduction through probabilistic extraction [15, 7]. A work is divided into a sequence of prefix-suffix window pairs (x, z). For a given prefix x, the model assigns a probability

$$p_z \;=\; p_\theta(z \mid x) \;=\; \prod_{t=1}^{|z|} p_\theta(z_t \mid x,\, z_{<t})$$

We call a window as *extractable* if $p_z \geq \tau$, we use $\tau = 0.001$ from [7].

Curated forget sets. Given a suppression request $\mathcal{W}_f$, target texts corresponding to $\mathcal{W}_f$, a trained model θ, and access to a large search corpus $\mathcal{C}$, a curator $\mathcal{A}$ returns a forget set D_f and, optionally, a retain set D_r. A downstream unlearning algorithm $\mathcal{U}$ then produces an unlearnt model

$$\theta' = \mathcal{U}(\theta, D_f, D_r).$$

In this paper, the curator $\mathcal{A}$ is the object under evaluation: we compare different choices of $\mathcal{A}$ while holding the downstream unlearning procedure fixed unless otherwise stated.

Search Corpora. For our analysis and experiments we use three distinct scale corpora, $\mathcal{C}_{\text{mid}}$ the `Dolmino` midtraining mix [32], $\mathcal{C}_{\text{pre}}$ a subset of the `Dolma3` pretraining mix [32], and $\mathcal{C}_{\text{CC25}}$ the Jan 2025 Common Crawl snapshot. See Appendix C for details.

Evaluation. The goal of verbatim output suppression is to make extractable windows from $\mathcal{W}_f$ un-extractable after unlearning, without inducing the same effect on extractable windows from $\mathcal{W}_r$. Verbatim suppression should not erase knowledge about a work, we also evaluate content-grounded question answering over the same works, together with general capability benchmarks. Thus, a curator is judged not by textual relevance alone, but by the behavior of the unlearnt model it induces. Figure 3 captures the CLEANSLATE pipeline. The central difficulty is that the documents in $\mathcal{C}$ that support a target continuation need not be canonical copies of the requested work. We next show that songs and books often have broad corpus footprints, making the mapping from $\mathcal{W}_f$ to D_f nontrivial.

4 The Corpus Footprint: Why Forget Set Curation Is Hard

The training evidence supporting a verbatim continuation need not be isolated to a canonical copy of the work. It may appear in lyric aggregators, forum discussions, fan fiction, or code snippets. A curator's true target is therefore a work's *corpus footprint*: the distributed set of documents and spans that can support the target continuation.

Measuring literal overlap. To quantify the corpus footprint of a work, we measure the scale of exact word-level overlaps between the target text and a corpus $\mathcal{C}$ using Infini-gram-mini [35]. For a given work, we compute localized n-gram occurrence counts across every position (see Appendix D for details). This is a conservative measure as changes in formatting, whitespace, or even case can break a match. It also misses paraphrases, translations, and semantic references, nor does it identify which documents actually caused a continuation. Instead, it lower bounds the literal overlap that a curator would miss if it searched only for canonical copies. At web scale, any large corpus will contain many short n-grams from almost any English text even when no canonical copy is present. So in addition we also measure the *coverage*: the fraction of the requested work covered by N-gram matches. This tracking reveals two primary mechanisms of data diffusion, unpacked below.

Outward diffusion. Works with cultural impact quickly spread after release. `Smile` (Morgan Wallen, released 31st Dec 2024) illustrates how rapidly a footprint forms: in $\mathcal{C}_{\text{CC25}}$ (Jan 2025), it exhibits 100% coverage at the 5-gram level, with some spans having $\geq 10^5$ occurrences, it also has 100% coverage at the 50-gram level with some spans having $\geq 10^4$ matches (see Figure 4). Clearly $\mathcal{C}_{\text{CC25}}$ contains canonical copies of `Smile`, but also a large cloud of shorter exact overlaps. Older works have had decades to diffuse, they are copied, quoted, remixed, discussed, and embedded in unexpected sources. Figure 1 illustrates this for *Never Gonna Give You Up*: beyond canonical lyric pages, the same text appears in fan transcriptions, wiki pages, code files, review prose, and question-answering sites. In $\mathcal{C}_{\text{mid}}$, this work has 94.2% coverage at $n = 5$ and 56.4% coverage at $n = 25$ (see Figure 5), with matched spans distributed across ≥ 147 unique documents. Any curator looking solely at canonical sources would miss the vast majority of these spans.

Inward diffusion. A new work may contain language that already existed in the corpus before the work itself was released. This can happen through idioms, genre templates, stock phrases, public-domain quotations, or repeated cultural language. Figure 2 shows several post-cutoff songs with substantial exact-overlap coverage in corpora that predate their release. For example, *Red Terror* by The Weeknd has roughly 20% coverage at $n = 7$ in $\mathcal{C}_{\text{mid}}$ because it quotes a 1910 poem. This shows that parts of the requested continuation may be supported by older or unrelated text. This makes curation more subtle than finding noisy and approximate copies of the target work: the relevant evidence may predate the work or appear in documents that do not look like copies at all.

Connection to target continuations. Corpus footprints help explain where verbatim output suppression becomes difficult. Figures 5–8 overlay localized corpus occurrence counts with extraction probabilities across songs, poems, and books. Across these case studies, extractable continuations often occur in regions with dense literal overlap, such as song choruses, famous quotations, or repeated phrases. We treat this as a diagnostic rather than a causal claim: high overlap does not prove that a particular document caused a completion, and low overlap does not rule out elicitation through other mechanisms. The practical implication is that a curator that misses dense regions of the footprint may leave the model enough data to be able to reproduce the requested continuation, while a curator that captures them too broadly may also affect neighboring content.

5 CLEANSLATE

Building on the formalism of Section 3, CLEANSLATE pairs each suppression request with per-model extractability evidence and content-grounded QA, so that the full pipeline of curator, unlearning algorithm, and edited model can be scored along several axes. Figure 3 summarizes the construction and evaluation flow.

Content domains. CLEANSLATE covers two domains, songs and books, whose corpus footprints (Section 4) are shaped differently. Songs travel as short, repetitive, discrete units. Books propagate

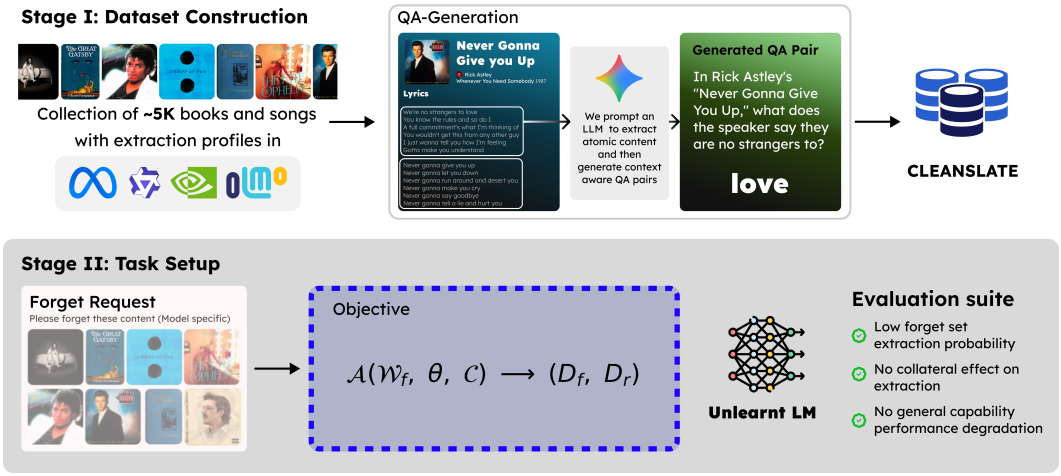


Figure 3: **CleanSlate overview.** *(Top)* We assemble ∼5K books and songs and measure baseline extraction profiles across different model families and generate QA pairs for each content. Together these form the CleanSlate dataset. *(Bottom)* Given a forget request $\mathcal{W}_f$ specifying k items, target model parameters θ, and an arbitrary corpus $\mathcal{C}$ (real, synthetic, or generated), the objective is to create a curator $\mathcal{A}$ that outputs a forget set $\mathcal{D}_f$ and retain set $\mathcal{D}_r$. These are passed to an unlearning procedure, and the resulting model is evaluated for low forget set extraction, preserved extraction on retained content, and stable general capability performance.

as longer-form passages reaching the corpus primarily through commentary and excerption rather than full copies. A curator that recovers one shape may miss most of the other, making the two domains complementary stress tests. We source songs from Billboard Hot 100 annual charts spanning 1970–2025, matched against the LRCLib lyrics database by title and artist, and 50 books, mostly from Project Gutenberg [26] with a small number of closed-license works.

Extraction profiles. We slide a window of 100 prefix and 100 suffix characters with stride 10 over each work, and label a work *model-extractable* for a base model θ if at least 5% of its windows are extractable in the sense of Section 3. The resulting profile is per-model. The same song may be extractable for one model and not for another. See Appendix E for more details.

Forget and retain pools. A suppression request is a model specific sample of size $|\mathcal{W}_f| = 50$ drawn from the model-extractable works of θ, with $\mathcal{W}_r$ defined as in Section 3. For a fixed model we reuse the same $\mathcal{W}_f$ and $\mathcal{W}_r$ across curators so that differences are attributable to the curation.

Content-grounded QA. Suppressing verbatim reproduction and erasing factual knowledge are distinct goals. A model that has been asked to stop reproducing a song's lyrics should still be able to answer factual questions about its content. We construct CleanSlate-QA in two stages. First, we prompt an LLM[1] to extract *atomic statements* from each work; factual statements anchored to a named entity, place, number, or concrete event mentioned inside the text, with priors such as title, creator, year, and genre explicitly excluded. Second, each statement is turned into a QA pair whose question embeds the title and creator naturally so it is self-contained, and whose answer is a short 1-5 word entity. For example, in J.K. Rowling's *Harry Potter and the Sorcerer's Stone* we ask where the Dursleys make Harry sleep, with answer *cupboard under the stairs*. The final dataset contains 12,088 QA pairs spanning the songs and books in $\mathcal{W}$.

End-to-end evaluation. For each curator $\mathcal{A}$, we measure baseline extractability and QA on θ, run $\mathcal{A}$ to obtain D_f, apply an unlearning algorithm $\mathcal{U}$ to obtain θ', and re-measure. $F \uparrow$ is the net percentage reduction in the number of extractable windows of works in $\mathcal{W}_f$ from θ to θ'; $R \downarrow$ is the same quantity over $\mathcal{W}_r$, where high values indicate collateral suppression. Negative values denote a net increase. *QA*

[1] `gemini-3.1-flash-lite-preview` via LiteLLM; Appendix F lists all models used.

Table 1: Curation comparison at $|\mathcal{W}_f| = 50$ with SimNPO unlearning, for three retrieval-based curators (BM25 pre, BM25 mid, Infini-gram mid) and the evaluation-aware curator (EA). F and R are net percentage reductions in extractable-window counts on the forget and retain pools; negative values denote net increases (Section 5). QA and validation columns report percentage-point changes from the corresponding model baselines (Table 4); *Avg* averages the six validation changes.

		Net reduction (%)		QA	Validation Δ (pp)						
	Model	F↑	R↓	Δ (pp)	GSM8K	BBH	WinoG	CoQA	HEval+	LMBDA	*Avg*
BM25 pre	GEMMA-3-12B	39.3%	38.4%	-2.3	+3.9	-4.7	+0.2	-6.4	+4.9	-4.9	-1.2
	OLMO-3-32B	43.1%	30.8%	-0.2	+1.5	+0.2	+0.6	+4.6	+0.0	+0.5	+1.2
	LLAMA-3.1-8B	28.9%	51.2%	-0.4	-0.1	+0.1	-0.6	+1.9	+1.2	-1.0	+0.3
	NEMOTRON-9B	14.6%	14.9%	+0.6	-11.1	-4.0	+0.9	+0.6	+0.6	-1.7	-2.4
	QWEN3-8B	1.6%	2.1%	-3.9	-1.4	+1.4	+3.2	-3.5	-6.7	+3.8	-0.5
	OLMO-3-7B	1.0%	2.9%	-0.1	+1.7	-0.1	-0.1	-3.8	-0.6	-0.0	-0.5
	Average	21.4%	23.4%	-1.1	-0.9	-1.2	+0.7	-1.1	-0.1	-0.5	-0.5
BM25 mid	GEMMA-3-12B	46.3%	46.0%	-3.2	+2.0	-8.6	-0.4	-9.2	+2.4	-6.3	-3.4
	OLMO-3-32B	43.0%	35.2%	+0.3	+2.4	+0.3	+1.4	+2.9	+0.6	+0.7	+1.4
	LLAMA-3.1-8B	16.7%	54.0%	-1.1	-0.2	-0.4	-1.5	-2.4	+1.8	-1.4	-0.7
	NEMOTRON-9B	13.0%	14.8%	+0.6	-10.1	-4.1	+0.6	-0.3	+0.6	-2.4	-2.6
	QWEN3-8B	2.3%	2.7%	-5.8	-1.1	+1.4	+3.6	-0.6	-10.4	+3.6	-0.6
	OLMO-3-7B	1.5%	3.7%	-0.0	+1.0	-0.2	+0.0	-1.9	-1.2	-0.3	-0.4
	Average	20.5%	26.1%	-1.5	-1.0	-2.0	+0.6	-1.9	-1.0	-1.0	-1.0
Infinigram mid	GEMMA-3-12B	47.8%	44.9%	-2.8	+3.2	-5.3	-0.2	-10.1	+1.8	-6.0	-2.8
	OLMO-3-32B	44.9%	36.0%	+0.4	+2.7	+0.1	+1.7	+3.9	+0.0	+0.9	+1.5
	LLAMA-3.1-8B	19.4%	55.7%	-0.8	+0.3	-0.4	-0.9	+0.9	+0.0	-1.4	-0.2
	NEMOTRON-9B	12.7%	14.5%	+0.5	-0.1	-2.3	+0.2	+0.3	+3.7	-2.3	-0.1
	QWEN3-8B	2.8%	2.8%	-4.6	-1.0	+1.5	+3.9	-3.1	-9.8	+3.6	-0.8
	OLMO-3-7B	1.6%	3.2%	-0.1	+0.9	-0.3	+0.0	-0.8	-0.6	-0.1	-0.2
	Average	21.6%	26.2%	-1.2	+1.0	-1.1	+0.8	-1.5	-0.8	-0.9	-0.4
EA	GEMMA-3-12B	96.2%	85.4%	-6.1	+3.6	-29.4	+0.0	-6.3	+1.2	-13.8	-7.4
	OLMO-3-32B	100.0%	97.0%	-2.7	+3.2	-3.3	-4.9	+3.9	-1.2	-3.0	-0.9
	LLAMA-3.1-8B	100.0%	100.0%	-5.7	-11.8	-37.6	-3.1	-3.9	+3.7	-10.1	-10.4
	NEMOTRON-9B	100.0%	80.3%	-1.2	-23.1	-22.8	+0.2	-5.5	-1.2	-5.2	-9.6
	QWEN3-8B	100.0%	82.7%	-21.9	-3.0	-2.0	+0.3	-7.9	-4.9	-6.0	-3.9
	OLMO-3-7B	100.0%	72.2%	-1.0	+2.8	-0.1	-2.0	-5.4	+1.8	-2.3	-0.9
	Average	99.4%	86.3%	-6.5	-4.7	-15.9	-1.6	-4.2	-0.1	-6.7	-5.5

Δ is the change in CleanSlate-QA accuracy. A validation suite covers math (GSM8K [6]), held-out reasoning (BBH [31]), commonsense (WinoGrande [29]), reading comprehension (CoQA [28]), code (HumanEval+ [21]), and language modeling (LAMBADA [27]). Per-model baselines are reported in Appendix E. CLEANSLATE supports two evaluation modes: request-level evaluation lets curator and unlearner vary jointly and scores the composed request-to-data-to-update pipeline; algorithm-focused evaluation holds two of ⟨curator, unlearner, model⟩ fixed and varies the third.

6 Experiments

We evaluate three retrieval curators and an evaluation-aware stress test using CLEANSLATE with $|\mathcal{W}_f| = 50$. Table 1 fixes the unlearner at SimNPO so row differences isolate the curator. Table 2 fixes $\mathcal{A}$ to the evaluation aware curator and varies the unlearner $\mathcal{U}$, (see Appendix G for hyperparameter details). Table 3 completes the curator × unlearner grid on three models. We perform experiments over six models, LLAMA-3.1-8B, OLMO-3-7B, NEMOTRON-9B, QWEN3-8B, GEMMA-3-12B, and OLMO-3-32B. Per-model pre-unlearning evaluations are in Appendix Table 4.

Curators. To evaluate whether standard corpus search can construct selective forget sets, we test two retrieval-based curators against large training corpora, alongside an evaluation-aware baseline: **BM25-pre** and **BM25-mid** use BM25 indices over chunked documents from $\mathcal{C}_{\text{pre}}$ and $\mathcal{C}_{\text{mid}}$, respectively.

Table 2: Algorithm comparison at $|\mathcal{W}_f| = 50$ with EA curation (SimNPO, UNDIAL, RMU). Metrics and columns follow Table 1.

		Net reduction (%)		QA	Validation Δ (pp)						
	Model	F↑	R↓	Δ (pp)	GSM8K	BBH	WinoG	CoQA	HEval+	LMBDA	*Avg*
SimNPO	LLAMA-3.1-8B	100.0%	100.0%	-5.7	-11.8	-37.6	-3.1	-3.9	+3.7	-10.1	-10.4
	OLMO-3-7B	100.0%	72.2%	-1.0	+2.8	-0.1	-2.0	-5.4	+1.8	-2.3	-0.9
	QWEN3-8B	100.0%	82.7%	-21.9	-3.0	-2.0	+0.3	-7.9	-4.9	-6.0	-3.9
	Average	100.0%	84.9%	-9.6	-4.0	-13.2	-1.6	-5.7	+0.2	-6.1	-5.1
UNDIAL	LLAMA-3.1-8B	100.0%	89.6%	-2.8	+1.4	-0.4	-0.3	-6.4	+3.7	-2.7	-0.8
	OLMO-3-7B	100.0%	79.6%	+0.2	-0.8	-0.9	-0.9	+11.4	+5.5	-0.3	+2.3
	QWEN3-8B	100.0%	83.1%	-5.2	-1.8	+0.7	+1.4	-5.4	-3.0	+0.9	-1.2
	Average	100.0%	84.1%	-2.6	-0.4	-0.2	+0.1	-0.1	+2.0	-0.7	+0.1
RMU	LLAMA-3.1-8B	98.8%	91.3%	-2.0	-10.7	-30.6	-1.6	-23.7	+1.2	-28.5	-15.6
	OLMO-3-7B	79.0%	81.4%	+0.6	+3.8	-9.0	+0.4	+8.2	+2.4	-14.7	-1.5
	QWEN3-8B	24.5%	25.7%	-2.3	-1.7	-5.6	-2.5	+1.6	+7.9	-3.8	-0.7
	Average	67.5%	66.1%	-1.2	-2.9	-15.0	-1.2	-4.6	+3.9	-15.7	-5.9

For each requested work, the curator keeps the top ranked retrieval units and converts them into completion examples. **Infini-gram-mid** uses an Infini-gram-mini index over $\mathcal{C}_{\text{mid}}$ to find maximal exact character spans of the requested text that occur in the corpus. The matched spans are then localized in retrieved corpus contexts. Finally, we test an Evaluation Aware **(EA)** curator which for each work in $\mathcal{W}_f$, constructs D_f to contain all the windows for that reference-text. EA presumes access to the evaluation windows and serves as a robustness test of the evaluation rather than a practical curator. All curators generate D_f in the same format: a 100-character prefix followed by a 100-character suffix. For BM25, we slide this window over the retrieved BM25 text units recorded in the retrieval artifact. For Infini-gram, we emit windows from the retrieved corpus context whose 200-character frame contains the midpoint of the exact match. This fixes sequence length but not the number of examples: the natural curator outputs are not size matched. We treat selected content and intervention volume as properties of the end-to-end curator output; with fixed epochs, output size also changes the number of optimizer updates. Exact output sizes are reported in Appendix K, and further retrieval details are in Appendix H.

6.1 Can natural retrieval curators induce selective suppression?

Under SimNPO, we find that retrieval engagement broadly tracks model size (Table 1). The two largest models (GEMMA-3-12B, OLMO-3-32B) show the strongest forget-side movement, LLAMA-3.1-8B and NEMOTRON-9B show partial movement, and the smallest two (OLMO-3-7B, QWEN3-8B) barely move. Retain-side movement follows the same ordering. Selectivity is model-determined rather than retriever-determined. LLAMA-3.1-8B loses between 51.2% and 55.7% of its retain-pool extractability across the three retrieval methods, while its forget side gains at most 28.9%. OLMO-3-32B is the only row where forget movement consistently exceeds retain movement. Switching from $\mathcal{C}_{\text{mid}}$ to $\mathcal{C}_{\text{pre}}$, or from BM25 to Infini-gram, reorders rows but does not change which models engage.

6.2 What if the target windows are given directly?

EA drives forget extractability to near-100% on every model, including the small models (OLMO-3-7B, QWEN3-8B) that retrieval barely moved (Table 1). This shows that the models can be moved by direct target-window interventions under SimNPO; the crossed study below shows that retrieval outcomes cannot be attributed to the curator, model, or unlearner in isolation. Retain-pool extractability also falls on every row. We hypothesize that this reflects corpus support shared across works (Section 4). Capability cost varies by an order of magnitude across the suite, with damage concentrated on BBH and LAMBADA.

Table 3: Curator × unlearner cross evaluation at $|\mathcal{W}_f| = 50$. Each forget set is reused across unlearners, so only the unlearner varies within a curator block. Entries average LLAMA-3.1-8B, OLMO-3-7B, and QWEN3-8B; per-model results are in Appendix J and Table 2. Metrics follow Table 1; these three-model means differ from its six-model averages.

		Net reduction (%)		QA	Validation Δ (pp)						
	Unlearner	F↑	R↓	Δ (pp)	GSM8K	BBH	WinoG	CoQA	HEval+	LMBDA	*Avg*
BM25 pre	SimNPO	10.5%	18.7%	-1.5	+0.1	+0.5	+0.8	-1.8	-2.0	+0.9	-0.2
	UNDIAL	98.2%	83.5%	-1.6	+0.3	-1.8	+1.1	-7.7	+1.4	-0.9	-1.3
	RMU	32.8%	35.5%	-2.3	-16.1	-21.7	-2.2	-16.4	+1.8	-21.4	-12.7
BM25 mid	SimNPO	6.8%	20.1%	-2.3	-0.1	+0.3	+0.7	-1.6	-3.3	+0.6	-0.6
	UNDIAL	92.9%	86.5%	-2.5	-3.0	-10.9	+1.0	-15.0	-1.4	-2.3	-5.3
	RMU	37.4%	48.2%	-11.3	-18.1	-32.1	-6.9	-23.3	-6.9	-35.7	-20.5
Infinigram mid	SimNPO	7.9%	20.6%	-1.8	+0.1	+0.3	+1.0	-1.0	-3.5	+0.7	-0.4
	UNDIAL	84.5%	79.0%	-2.1	+0.2	-4.1	+1.1	-8.5	+1.2	-2.1	-2.0
	RMU	34.8%	44.3%	-7.3	-17.6	-28.5	-4.5	-19.8	-1.4	-33.8	-17.6
EA	SimNPO	100.0%	84.9%	-9.6	-4.0	-13.2	-1.6	-5.7	+0.2	-6.1	-5.1
	UNDIAL	100.0%	84.1%	-2.6	-0.4	-0.2	+0.1	-0.1	+2.0	-0.7	+0.1
	RMU	67.5%	66.1%	-1.2	-2.9	-15.0	-1.2	-4.6	+3.9	-15.7	-5.9

6.3 Are evaluation aware curator's effects algorithm-specific?

To determine if the EA curator's collateral damage is specific to SimNPO, we evaluate two additional unlearning algorithms (UNDIAL and RMU) on LLAMA-3.1-8B, OLMO-3-7B, and QWEN3-8B (Table 2). Across SimNPO and UNDIAL, F reaches 100% on all three models, and R stays between 72–100%. While both algorithms suffer from severe collateral forgetting by degrading retain-pool extractability, they produce very different capability outcomes on identical EA inputs. UNDIAL largely preserves average validation accuracy ($+0.1$ pp) while SimNPO degrades it (averaging -5.1 pp). RMU reduces F on LLAMA-3.1-8B and OLMO-3-7B, with retain extractability falling alongside. On QWEN3-8B, both forget and retain stall together at F=24.5% and R=25.7%, consistent with the hyperparameter sensitivity reported by Li et al. [18]. The same retain-pool damage and capability gap persist at $|\mathcal{W}_f| = 100$ on LLAMA-3.1-8B and QWEN3-8B. The full table is in Appendix I. Table 3 repeats this comparison for the three retrieval curators, reusing an identical D_f within each ⟨model, curator⟩ pair so that only the unlearner changes. Holding the curator fixed, the unlearner changes the outcome dramatically: the same BM25-pre forget sets yield 10.5% average forgetting under SimNPO, 98.2% under UNDIAL, and 32.8% under RMU, so weak suppression under SimNPO is not an intrinsic property of retrieval-derived forget sets. No configuration is selective, however: UNDIAL's strong suppression comes with 79.0–86.5% retain-side suppression, per-model results in Appendix J.

Takeaway. Retrieval-derived forget sets compose unpredictably with the downstream unlearner. Fixed-D_f algorithm comparisons remain valid, but their conclusions are conditional on the upstream curator; request-level comparisons evaluate the composed curator–unlearner pipeline. The evaluation-aware EA achieves forgetting on every model under SimNPO and UNDIAL, but the retain side falls and capability drops by up to 10.4 pp on average. At the tested operating points, neither approach is selective. Lexical retrieval underspecifies the corpus support of a work because target text is widely shared across documents that are not the canonical source. Target-window indexing overspecifies it because the extractable windows for one work might be correlated with those for others. The same EA inputs produce a 5-pp capability gap between SimNPO and UNDIAL, reinforcing that selective unlearning requires co-designing the curation strategy alongside the unlearning objective.

7 Discussion and Future Direction

Unlearning needs curation. Our results suggest that the forget set should not be treated as a fixed premise of language-model unlearning. In realistic deployments, the request is often stated at the level of a work or behavior, while the unlearning algorithm requires concrete data to update against.

What is selected for forgetting determines both whether the requested continuation becomes difficult to elicit and what else is disturbed. Forget set curation is therefore part of the unlearning problem, not merely a preprocessing detail.

Two insufficient endpoints. The experiments expose two natural but incomplete approaches to curation. Off-the-shelf corpus retrieval is not selective: lexical and exact-substring search can recover pieces of a work's corpus footprint, but the resulting forget sets compose unpredictably with the downstream unlearner and do not induce selective verbatim suppression at the tested operating points. Conversely, evaluation aware curation removes the retrieval bottleneck by selecting target windows directly, but still causes substantial output suppression of non-requested continuations and model-dependent capability regressions. Thus, the problem is not simply to find more target-like text. It is to construct a forget set whose effects remain localized after the model update.

Toward algorithm-aware curation. This localization depends on the downstream unlearning algorithm. The same evaluation aware forget set produces different capability profiles under SimNPO, UNDIAL, and RMU, suggesting that curation and unlearning should be evaluated jointly rather than as independent modules. Future curators may need to combine corpus-footprint signals with model-specific extraction profiles, dense or hybrid retrieval, influence estimation, or datamodeling to predict which examples will suppress the requested behavior without unnecessary collateral effects. CLEANSLATE studies a deliberately narrow setting: verbatim output suppression for songs and books. It does not address all forms of concept, entity, or factual unlearning, and our exact-overlap diagnostics miss paraphrase, translation, and semantic reuse. These limitations are also what make the task measurable. Extending forget set curation to richer request types, broader corpora, and algorithm-aware selection is a natural next step. The broader message is that practical unlearning should be studied end-to-end, from request, to curated forget set, to edited model.

Acknowledgments

We thank Ludwig Schmidt, Etash Guha, Pratyush Maini, and Ananjan Nandi for helpful discussions. We acknowledge compute support by the Center for AI Safety. SK acknowledges support by NSF 2046795 and 2205329, IES R305C240046, ARPA-H, the MacArthur Foundation, Schmidt Sciences, HAI, OpenAI, Microsoft, and Google.

References

[1] Ahmed Ahmed, A Feder Cooper, Sanmi Koyejo, and Percy Liang. Extracting books from production language models. *arXiv preprint arXiv:2601.02671*, 2026.

[2] Youssef Allouah, Rachid Guerraoui, and Sanmi Koyejo. Distributional machine unlearning via selective data removal. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IPqUBL4R9x.

[3] Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, Jun Zhao, et al. Rwku: Benchmarking real-world knowledge unlearning for large language models. *Advances in Neural Information Processing Systems*, 37:98213–98263, 2024.

[4] 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. 2021. URL https://arxiv.org/abs/2012.07805.

[5] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. 2023. URL https://arxiv.org/abs/2202.07646.

[6] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[7] A. Feder Cooper, Aaron Gokaslan, Ahmed Ahmed, Amy B. Cyphert, Christopher De Sa, Mark A. Lemley, Daniel E. Ho, and Percy Liang. Extracting memorized pieces of (copyrighted) books from open-weight language models. 2025. URL https://arxiv.org/abs/2505.12546.

[8] Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. Undial: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8827–8840, 2025.

[9] Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. 2025. URL https://arxiv.org/abs/2506.12618.

[10] Logan Engstrom, Axel Feldmann, and Aleksander Madry. DsDm: Model-aware dataset selection with datamodels. *arXiv preprint arXiv:2401.12926*, 2024.

[11] Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning, 2025. URL https://arxiv.org/abs/2410.07163.

[12] Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing conceptual knowledge from language models. *arXiv preprint arXiv:2410.02760*, 2024.

[13] Kristian Georgiev, Roy Rinberg, Sung Min Park, Shivam Garg, Andrew Ilyas, Aleksander Madry, and Seth Neel. Attribute-to-delete: Machine unlearning via datamodel matching. *arXiv preprint arXiv:2410.23232*, 2024.

[14] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

[15] Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, Ilia Shumailov, Milad Nasr, Christopher A. Choquette-Choo, Katherine Lee, and A. Feder Cooper. Measuring memorization in language models via probabilistic extraction. 2025. URL https://arxiv.org/abs/2410.19482.

[16] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

[17] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[18] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

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

[20] Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, et al. Olmotrace: Tracing language model outputs back to trillions of training tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 178–188, 2025.

[21] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation, 2023. URL https://arxiv.org/abs/2305.01210.

[22] Ken Ziyu Liu, Christopher A. Choquette-Choo, Matthew Jagielski, Peter Kairouz, Sanmi Koyejo, Percy Liang, and Nicolas Papernot. Language models may verbatim complete text they were not explicitly trained on, 2025. URL https://arxiv.org/abs/2503.17514.

[23] Ziwen Liu, Huawei Lin, Yide Ran, Denghui Zhang, Jianwen Xie, Chuan Li, Weijie Zhao, and Zhaozhuo Xu. Randomized antipodal search done right for data pareto improvement of llm unlearning. *arXiv preprint arXiv:2604.16591*, 2026.

[24] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. Tofu: A task of fictitious unlearning for llms. 2024. URL https://arxiv.org/abs/2401.06121.

[25] Soumyadeep Pal, Changsheng Wang, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. Llm unlearning reveals a stronger-than-expected coreset effect in current benchmarks. *arXiv preprint arXiv:2504.10185*, 2025.

[26] Project Gutenberg. Project gutenberg. https://www.gutenberg.org, 1971. Accessed March 11, 2026.

[27] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[28] Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question answering challenge, 2019. URL https://arxiv.org/abs/1808.07042.

[29] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

[30] 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. 2024. URL https://arxiv.org/abs/2407.06460.

[31] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022. URL https://arxiv.org/abs/2210.09261.

[32] Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

[33] Yixin Wan, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Rahul Gupta. Not every token needs forgetting: Selective unlearning to limit change in utility in large language model unlearning. *arXiv preprint arXiv:2506.00876*, pages 622–632, 2025.

[34] Ruihan Wu, Chhavi Yadav, Russ Salakhutdinov, and Kamalika Chaudhuri. Evaluating deep unlearning in large language models. *arXiv preprint arXiv:2410.15153*, 2024.

[35] Hao Xu, Jiacheng Liu, Yejin Choi, Noah A Smith, and Hannaneh Hajishirzi. Infini-gram mini: Exact n-gram search at the internet scale with fm-index. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24955–24980, 2025.

[36] Yuanshun Yao and Xiaojun Xu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

[37] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

[38] Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Douglas Zytko, Prashant Khanduri, and Dongxiao Zhu. Not all tokens are meant to be forgotten, 2025. URL https://arxiv.org/abs/2506.03142.

[39] Xiaoyuan Zhu, Muru Zhang, Ollie Liu, Robin Jia, and Willie Neiswanger. Llm unlearning without an expert curated dataset. In *Second Conference on Language Modeling*, 2025.

Appendix

A Additional Related Work

Unlearning algorithms and benchmarks. A range of algorithms have been proposed for language-model unlearning. Gradient ascent on the forget loss is the simplest baseline. Negative preference optimization and SimNPO [37, 11] cast unlearning as preference optimization against a reference model. Representation Misdirection Unlearning [18] perturbs internal representations on forget data while regularizing retain representations, and UNDIAL [8] adjusts logits through self-distillation. Several benchmarks evaluate whether such methods can remove specified behaviors while preserving utility. TOFU [24] provides synthetic author biographies, MUSE [30] evaluates multiple aspects of language-model unlearning including verbatim memorization, WMDP [18] evaluates hazardous-knowledge unlearning, and OpenUnlearning [9] provides a unified benchmarking framework. These benchmarks differ in domain and objective, but they largely evaluate the downstream unlearning step after the forget data, target examples, or target behaviors have already been specified.

Request-level, entity-level, and fact-level unlearning. Some recent work moves closer to settings where the target is specified at a higher level than a fixed forget corpus. RWKU [3] studies real-world knowledge unlearning, where the algorithm receives a target entity and the original model rather than an explicit training corpus. It then uses synthetic data produced by the model under evaluation to construct a forget set. This is close in spirit to our setting because the forget set is not directly provided, but the evaluated artifact differs: RWKU primarily measures unlearning algorithms under synthetic forget sets, while we evaluate the curation of forget sets under fixed unlearning algorithms. Deep fact unlearning [34] studies whether a target fact remains inferable from retained facts under logical rules. This is related to our setting in that a target behavior can be supported by non-target evidence, but the unit and objective are different: deep unlearning targets factual deductive closure, while we study verbatim continuation and ask which corpus spans or documents should be used for suppression.

Data for unlearning. Recent work shows that the contents of a forget set matter even after the forget corpus has been specified. [25] find that small subsets of benchmark-provided forget sets can match full-set unlearning. [33] and [38] show that token-level selection within known forget examples can reduce utility loss. [2] formalize distributional unlearning as selective data removal: given identified unwanted and retained sample sets, choose a small subset whose removal moves the edited data distribution away from the unwanted domain while preserving the retained one. These works are closely related to ours in emphasizing that selection matters, but they assume that the unwanted samples, retained samples, or target domain have already been obtained through an upstream process such as filtering, classification, or annotation. We study that upstream process itself. Zhu et al. [39] identify the construction of forget sets as a major bottleneck in unlearning pipelines, and synthesize proxy forget data given only a broad domain name as input. This is close to our setting because it addresses the upstream dataset construction problem, but differs in focus, they work on synthesizing proxy data for broad domain-level forgetting, while we focus on measuring the role of the forget set (and the distributed corpus footprint) for suppressing verbatim output of specific targeted individual works.

B Compute Requirements

All experiments were conducted on a compute node equipped with 8 NVIDIA H200 GPUs (141GB VRAM each), 230 CPU cores, 3TB of RAM, and 60TB of local NVMe storage. While model training utilized 1–2 GPUs, evaluation and validation tasks were performed on a single GPU. Notably, the corpus search and forget set curation phases are significantly memory- and storage-bound due to the scale of the datasets involved; these stages necessitated the full utilization of the available system memory and high-speed disk I/O.

C Search Corpora ($\mathcal{C}$) Details

For our experiments we use three distinct scale corpora

1. $\mathcal{C}_{\text{mid}}$ **(Midtraining):** `allenai/dolma3_dolmino_mix-10B-1025` [32] this is a midtraining mix which has coverage of categories like instruction tuning data (FLAN, Tulu-3-SFT), code (`cranecode`, `stack_edu`), and synthetic reasoning traces (Gemini, QwQ, Llama Nemotron, OpenThoughts).
2. $\mathcal{C}_{\text{pre}}$ **(Pretraining):** A $\sim 11\%$ subset of Dolma-3 6T (`allenai/dolma3_mix-6T`) pretraining mix [32]. To make it computationally feasible, we limit this subset to five Common Crawl shards of the full pretraining corpus, `art_and_design`, `entertainment`, `history_and_geography`, `literature`, and `religion`.
3. $\mathcal{C}_{\text{CC25}}$ **(Web Scale):** January 2025 Common Crawl snapshot with $\sim 9T$ tokens, accessed via the index provided by Infini-gram-mini [35]. Used only for analysis of corpus footprint in Section 4

Team Olmo et al. [32] states December 2024 as the knowledge cutoff for $\mathcal{C}_{\text{mid}}$ and $\mathcal{C}_{\text{pre}}$.

D Computing Coverage of Songs and Books in Large Corpus

We compute these statistics using Infini-gram-mini [35] but one can use any other dataset search tool.

D.1 Computing $N-$Gram Matches and Coverage

To quantify the verbatim overlap between a reference sequence and a large-scale corpus $\mathcal{C}$, we employ an iterative retrieval algorithm that identifies the maximal exact n-gram matches starting at every word start position. Let $T_w = (x_1, \ldots, x_m)$ be the whitespace-delimited word sequence of work w. For start position i and length n, let $s_{i,n}$ be the exact character span covering $x_i, \ldots, x_{i+n-1}$, preserving punctuation and spacing. For indexed corpus $\mathcal{C}$, we define the frequency $C_{\mathcal{C}}(s)$ and the coverage statistic $\text{cov}_{\geq N}(w; \mathcal{C})$ as follows:

$$\begin{aligned} C_{\mathcal{C}}(s) &= \sum_{d \in \mathcal{C}} \#_d(s) \\ \text{cov}_{\geq N}(w; \mathcal{C}) &= \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}[\,\exists\, i, n :\ n \geq N,\ i \leq j < i+n,\ C_{\mathcal{C}}(s_{i,n}) > 0\,] \end{aligned} \tag{1}$$

The statistic $\text{cov}_{\geq N}$ measures the fraction of word positions in w covered by at least one verbatim n-gram of length $n \geq N$ present in $\mathcal{C}$. For each start position i, we begin at $n_{\min} = 5$ and extend n while $C_{\mathcal{C}}(s_{i,n}) > 0$. Short n-grams ($n < 5$) occur with high background frequency due to linguistic coincidence.

D.2 Aggregate Coverage Statistics

Out of 4,663 works in CLEANSLATE , 4,596 (98.6%) retrieve at least one positive-count 5-gram match with documents in $\mathcal{C}_{\text{mid}}$. The distribution of retrieved documents per work exhibits a median of 119, a 90th percentile of 299, and a maximum of 45,338. For source inspection in $\mathcal{C}_{\text{mid}}$ we sample $D = 20$ documents for each of the top-K ($K = 20000$) spans ranked by length and occurrence count, for $\mathcal{C}_{\text{CC25}}$ we sample $D = 2$ documents for each of the top $K = 2000$ spans, due to computational limitations.

D.3 Extractability vs. Footprint Density Plots

To understand the correlation between localized corpus prevalence and model extractability, we plot the *footprint density* at a given word position j, alongside the maximum extraction probability (p_z) measured amongst all suffixes z covering the position j. The density aggregates the occurrence counts of all valid n-grams in $\mathcal{C}_{\text{mid}}$ that overlap the j-th word of the work.

Figures 5 through 8 corroborate that peaks in $\mathcal{C}_{\text{mid}}$ occurrences strongly align with spikes in extractability. For instance, the choruses of popular songs (*Never Gonna Give You Up*, *Rocket Man*)

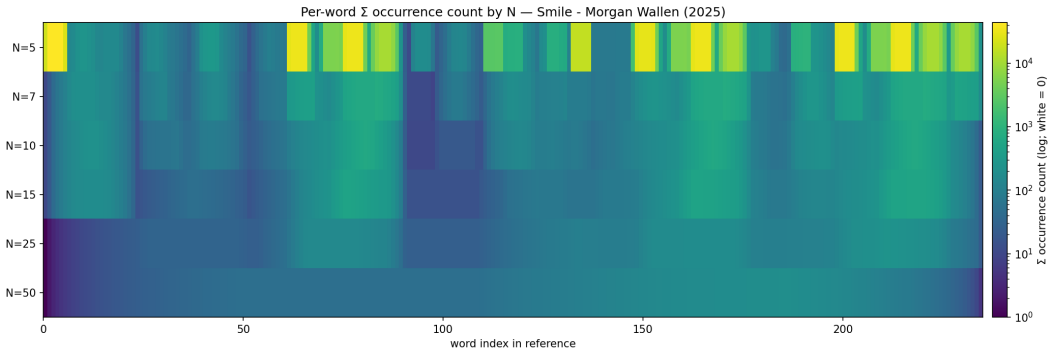


Figure 4: N-gram coverage for *Smile* in $\mathcal{C}_{\mathrm{CC25}}$

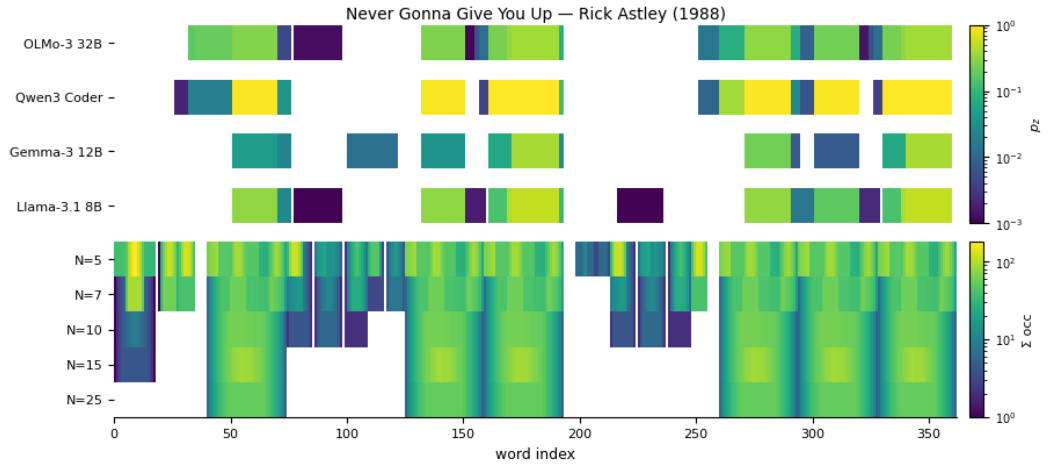


Figure 5: Footprint density (sum of matches of all overlapping n-gram matches in $\mathcal{C}_{\mathrm{mid}}$) vs. maximum extraction probability (p_z) per word for *Never Gonna Give You Up*.

and famous refrains in poems (*A Dream Within a Dream*) exhibit massive frequency spikes in the corpus. The model's extraction probability neatly mirrors these spikes, rising precisely where the footprint density is highest.

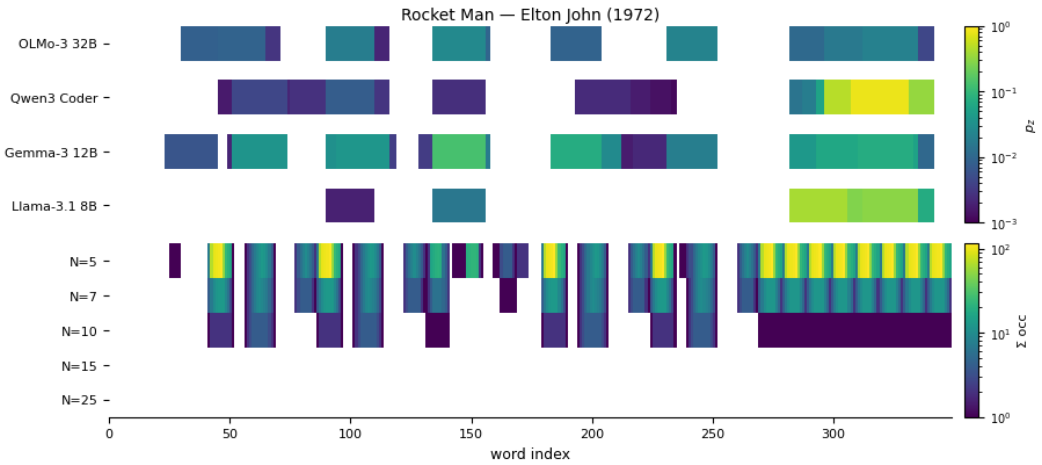


Figure 6: Footprint density vs. maximum extraction probability (p_z) per position for *Rocket Man*.

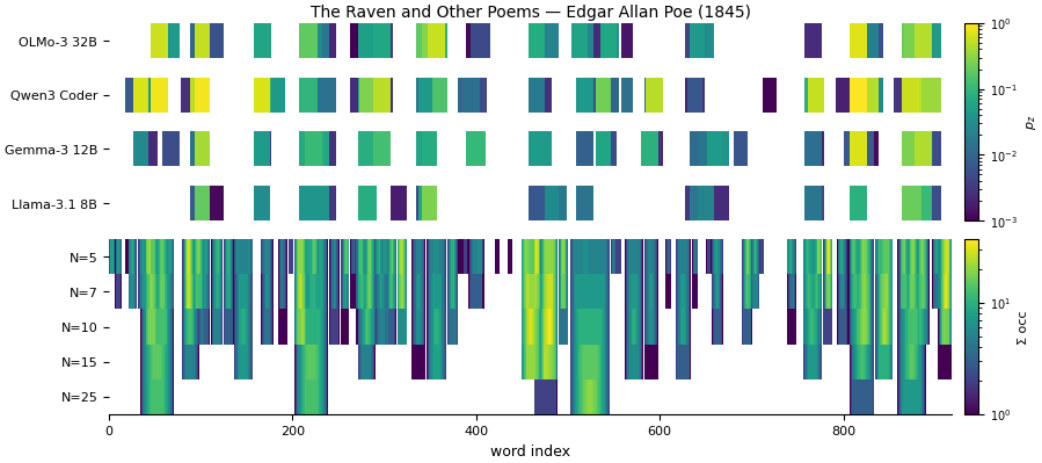


Figure 7: Footprint density in $\mathcal{C}_{\text{mid}}$ vs. maximum extraction probability (p_z) per position for poems of Edgar Allan Poe

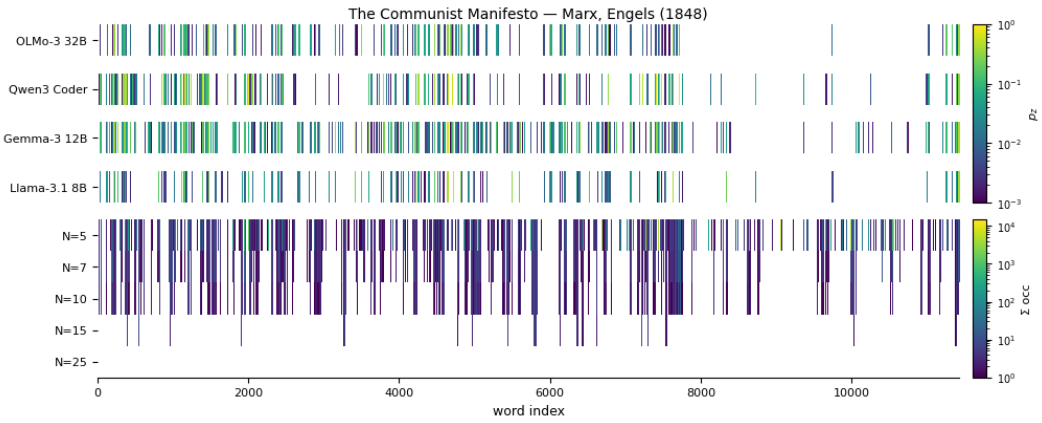


Figure 8: Footprint density in $\mathcal{C}_{\text{mid}}$ vs. maxp_z per position for *The Communist Manifesto*.

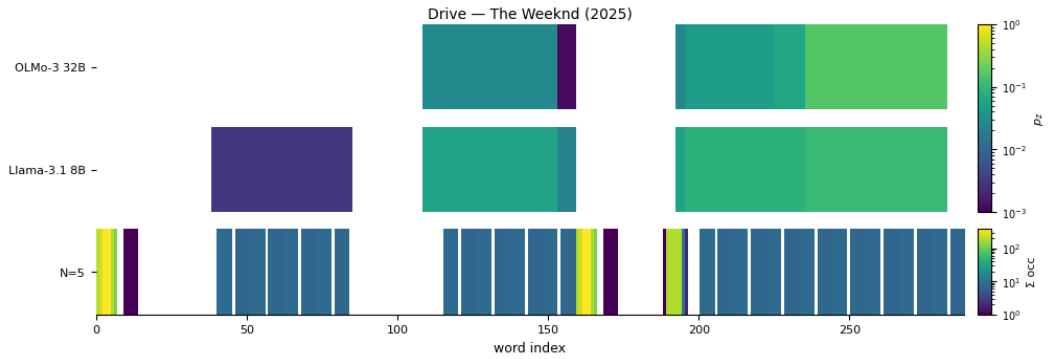


Figure 9: $\max p_z$ across *Drive* by Weekend, both models pre date the song, windows with high p_z are highly repetitive portions of the chorus

E Baseline Extractability Patterns Across Model Families

Extractability is neither uniform across target works nor consistent across model architectures. Table 4 documents the baseline characteristics of our evaluation suite prior to any curation or unlearning. The extraction threshold ($p_z \geq 0.001$) is evaluated over sliding windows of 100 prefix and 100 suffix characters.

Model scale and capability. Larger models consistently exhibit higher raw extractability. For instance, the 32B-parameter OLMO-3-32B achieves a forget-pool extraction rate (Ext-F) of 8.27%, while its 7B-parameter counterpart (OLMO-3-7B) achieves 7.20%. Base models also tend to exhibit higher extractability than their instruction-tuned variants, probably due to alignment training penalizing raw regurgitation in favor of conversational formatting. Repetitive structure of certain songs also increase their extractability, *Drive* by The Weeknd (2025) post dates Llama-3.1-8B and Olmo-32b but both model assign high p_z where a sections of the work are near identical prefix–suffix pair (see Figure 9), this suggests that the language-modeling objective by itself can increase the probability of extraction for highly repetitive works.

Heterogeneity of extractable text. The data reveals that extractability is highly localized within the texts themselves. For songs, verses often fall below the extraction threshold, while choruses reinforced by both their internal repetition and their higher external footprint density cross the threshold easily. More importantly, *different models extract different windows*. A specific verse that is highly extractable for LLAMA-3.1-8B may fall below the threshold for NEMOTRON-9B due to differences in their respective training mixtures.

QA performance vs. verbatim extraction. Table 4 also reports the baseline CLEANSLATE-QA accuracy (ranging from 22.01% to 35.07% pass@5). The fact that these models can reliably answer factual questions about the content confirms that they possess abstract knowledge of the works. The core challenge of forget set curation is to selectively suppress the localized p_z spikes responsible for exact extraction while leaving this broader semantic knowledge (QA) and general capabilities (GSM8K, etc.) undisturbed.

F CleanSlate-QA Benchmark Construction

Here we go over the pipeline used to build CleanSlate-QA, the content-grounded retain-metric benchmark referenced in Section 5 and reported as the *QA* column of Table 4. Construction has three stages. Stage 1 extracts factual propositions from each work, stage 2 turns each proposition into an atomic question-answer pair, and stage 3 filters candidates by out-of-sample (OOS) model knowledge.

Table 4: Baseline evaluation models prior to any curation or unlearning, ordered by forget-pool extractability. The *Forget pool* and *Retain pool* blocks report the total number of $(100, 100)$ character windows (N) and the number that meet the extraction threshold $p_z \geq 0.001$ (*Ext.*). *QA* is CleanSlate-QA pass@5 accuracy (%) averaged over forget and retain items. Remaining columns report baseline accuracy (%) on the validation suite. Forget pool size is $|\mathcal{F}|=50$ except for Olmo-3-7B, which has $|\mathcal{F}|=38$. GSM8K uses flexible-extract scoring.

	Forget pool		Retain pool		QA (%)	Validation accuracy (%)						
Model	N	Ext.	N	Ext.		GSM8K	BBH	HSwag	WinoG	CoQA	HEval+	LMBDA
Llama-3.1-8B	8,513	599	3,999,813	11,130	25.63	50.64	64.57	60.69	74.43	49.10	31.71	76.29
Olmo-3-7B	8,084	582	4,000,242	2,108	27.31	69.60	61.63	54.13	69.77	57.63	31.71	69.63
Nemotron-9B	8,840	647	3,999,486	3,082	22.01	76.80	72.67	58.11	73.48	53.92	10.98	68.00
Qwen3-8B	10,128	755	3,998,198	3,297	34.71	88.55	79.48	57.12	67.88	64.78	58.54	65.17
Gemma-3-12B	27,936	2,210	3,980,390	14,501	30.63	71.80	73.91	62.04	75.30	50.03	12.80	76.03
Olmo-3-32B	67,323	5,565	3,941,003	17,280	35.07	78.24	78.73	60.97	76.48	58.38	39.63	76.31

Routing and infrastructure. Each work in $\mathcal{W}$ is routed by length. Books are chunked into 60,000-character windows that are extracted while songs are extracted in a single call. Stages 1 and 2 dispatch `gemini-3.1-flash-lite-preview`. Stage 3 uses (`Qwen/Qwen3.5-9B`, `meta-llama/Meta-Llama-3-8B-Instruct-Lite`).

F.1 Stage 1: Proposition Extraction

The model is asked to extract a small set of *interior factual propositions*, defined as statements anchored to a named entity, place, number, or concrete event mentioned inside the text, with priors such as title, creator, year, and genre explicitly excluded. Each proposition must cite a short source span from the work. The song and book prompts are reproduced below.

Song Proposition Extraction Prompt

You are extracting interior facts from a song's lyrics that will be used to build a retain-metric QA benchmark.

Title: "`{title}`"
Creator: `{creator}`

LYRICS:
`{text}`

TASK: Extract between 3 and 8 **interior factual propositions** from these lyrics.

A good proposition:

- Refers to a named entity, specific detail, number, place, person, action, or relationship mentioned INSIDE the lyrics.
- Is a full factual statement, not just a word.
- Cites a short source span (1–2 lines from the lyrics) where the fact appears.

A bad proposition:

- Is about the title, creator/artist, year, genre (priors).
- Restates the hook, chorus, or title in different words (e.g. for "Ladies' Night": "the song says it is your night" — that's just the hook).
- Has no named entity, place, number, or specific concrete detail.
- Is a generic feeling, mood, theme, or exhortation ("the song tells listeners to dance").
- Uses pronouns/possessives as its key content ("your", "their", "this").

If the song is abstract with few concrete details, return fewer propositions. Do NOT pad.

Return ONLY a JSON object:
`{"propositions": [{"fact": "...", "source_span": "..."}, ...]}`

Book Proposition Extraction Prompt (per chunk)

You are extracting interior facts from a CHUNK of a book that will be used to build a retain-metric QA benchmark.

Book: "`{title}`"
Author: `{creator}`
Chunk `{chunk_idx}` of `{total_chunks}`

TEXT:
`{text}`

TASK: Extract between 5 and 15 **interior factual propositions** from this chunk.

Prioritize:

- Named characters and their distinguishing features, relationships, actions.
- Specific locations, settings, objects mentioned.
- Concrete plot events in this chunk.
- Numeric specifics (ages, dates within the narrative, counts).

Avoid:

- Anything about the title, author, publication year, genre (priors).
- Generic theme/mood statements without a concrete anchor.
- Verbatim famous quotes unless they reveal a concrete fact.

Each proposition must cite a short source span (1–3 lines) from the chunk.

Return ONLY a JSON object:
`{"propositions": [{"fact": "...", "source_span": "..."}, ...]}`

F.2 Stage 2: Atomic QA Generation

Each proposition is converted into one atomic question-answer pair. Questions must be self-contained (naming both title and creator), target an interior detail rather than a prior, and admit a short 1–5 word answer that is itself a named entity, number, specific place, specific object, or proper noun. The model is allowed to return an empty list when no valid pair can be produced for a given proposition. This is the primary mechanism by which low-yield propositions are dropped.

Atomic QA Generation Prompt

Turn this fact into ONE atomic question–answer pair for a retain-metric benchmark.

Title: "`{title}`"
Creator: `{creator}`
Fact: `{fact}`
Source span: `{source_span}`

REQUIREMENTS — the question must:

1. Be SELF-CONTAINED (atomic): name the title AND creator naturally. The model sees only the question, no context.
2. Have a SHORT semantic answer (1–5 words) that is a **named entity, number, specific place, specific object, or proper noun**. Not a pronoun, possessive, or generic colloquial phrase.
3. Target an interior detail — NOT the title, creator, year, or genre.
4. The answer must NOT be the title or any substring of the title/creator.
5. The answer must NOT appear as a phrase inside your own question text.
6. Not be answerable from the title alone (e.g., if the song is "Sexy + 17" don't ask the girl's age).

GOOD examples:

- Q: "In Dr. Hook's 'Sylvia's Mother', how does the caller address Sylvia's mother?" A: "Mrs. Avery"

Table 5: Unlearning hyperparameters.

	SimNPO	UNDIAL	RMU
β / δ	2.0 / 0.0	–	–
β_{U}	–	10.0	–
RMU coeff. / layer	–	–	2.0 / `layers.7`
λ_f / λ_r	0.125 / 1.0	1.0 / 1.0	1.0 / 1.0
Retain loss	NLL	NLL	embed-diff
Epochs	1	2	2
LR / warmup	10^{-5} / 10	10^{-5} / 20	10^{-5} / 20
Batch / accum.	8 / 4	8 / 4	8 / 4

- Q: "In 'C'mon N' Ride It' by Quad City DJ's, what car does the narrator want to be in the back of?" A: "Impala"

BAD examples:

- Q: "Who sang 'Hotline Bling'?" → priors
- Q: "Complete: 'Ride that choo-choo ___' " → verbatim
- Q: "What is the mood of 'More Than A Woman'?" → subjective
- Q: "In Kool & the Gang's 'Ladies' Night', what does the song say is happening tonight?" A: "your night" → hook-echo, pronoun answer, no specific content

If the fact cannot produce a good atomic Q meeting all requirements, return an empty list.

Return ONLY a JSON object shaped like:
`{"pairs": [{"question": "...", "answer": "..."}, ...]}`

F.3 Stage 3: OOS Knowledge Filter and Judge

To drop questions whose answers can only be recovered by memorizing the source work, each candidate question is presented to the OOS probe models with no surrounding context. A candidate is kept iff at least one probe model produces a correct response.

LLM Judge Prompt

Does the model answer contain the reference answer or convey the same meaning?

Reference: `{reference}`
Model Answer: `{model_answer}`

Reply with ONLY this JSON, nothing else: `{"correct": true}` or `{"correct": false}`

G Unlearning Training Details

The trainer consumes curated forget rows of the form (x, z), where x is a prefix and z is the suffix to suppress. Each training item is anchored on one forget example and paired with a randomly sampled retain example. The total loss is

$$\lambda_f \mathcal{L}_{\mathrm{forget}} + \lambda_r \mathcal{L}_{\mathrm{retain}}.$$

For SimNPO we use the average suffix NLL form

$$\mathcal{L}_{\mathrm{SimNPO}} = -\frac{2}{\beta} \log \sigma\big(\beta(\bar{\ell}_\theta(z \mid x) - \delta)\big)\,,$$

with $\mathcal{L}_{\mathrm{retain}}$ equal to retain NLL. UNDIAL distills from a frozen reference model after subtracting β_{U} from the gold-token teacher logit on forget suffix tokens. RMU minimizes MSE between forget activations and a random control vector at decoder block `model.layers.7`, and uses an activation matching retain loss against the frozen reference model. Table 5 captures all hyperparameters specific to $\mathcal{U}$.



Figure 10: Example QA pairs from CLEANSLATE

G.1 Some example QA pairs

See Figure 10

H Retrieval Curators and Forget-Set Construction Details

Each retrieval curator maps a request text to a set of prefix–suffix training examples in three steps: retrieve corpus units, project retrieved text into evaluation-shaped windows, and allocate a bounded

number of windows per requested work. The unlearning trainer consumes only these pre-sliced rows: it tokenizes the concatenated prefix and suffix and masks prefix tokens from the forget loss.

BM25 retrieval. For BM25, we build sharded BM25s indices over the search corpus. Documents shorter than 100 characters are discarded. Remaining documents are split at word boundaries into segments of at most 2,000 characters with 500-character overlap; segments shorter than 200 characters are removed. BM25 tokenization lowercases text and removes English stopwords. At search time, each shard returns its top candidates and we retain the global top 100 segments for the requested work. Very long request texts are queried through sampled probes, after which the same global top-k merge is applied.

Infini-gram retrieval. We use an Infini-gram-mini over the corpus. For each character offset i in the request text T, we first test whether the 20-character substring beginning at i occurs in the corpus. If it does, similar to OlmoTrace [20] we binary search for the longest substring $T[i : j]$ with positive corpus count and record the span together with its count. After scanning all offsets, we rank the spans by character length, and keep the top 100. For each retained span, the retrieval pipeline queries corpus contexts containing the span; the curation artifact stores the first retrieved context, while metadata records the matched `span_text`, corpus count, coverage, and number of retrieved contexts.

Window extraction. We use the same character window shape for all forget sets. Let $W = 100$ and $S = 10$. A window beginning at character position s emits

$$\text{prefix} = D[s : s + W], \qquad \text{suffix} = D[s + W : s + 2W].$$

EA curator applies this sliding window directly to the reference text. BM25 applies it to the retrieved BM25 text unit recorded in the retrieval artifact. Infini-gram based curator first locates `span_text` inside the retrieved corpus context and emits stride-spaced windows whose 200-character frame contains the midpoint of the matched span. Duplicate prefix- suffix pairs are removed within each requested work.

Quota allocation. Retrieval can produce many overlapping windows from a small number of documents. To avoid letting a single repeated source dominate the forget set, we cap each requested work at $C = 128$ windows and allocate this budget across retrieved documents. Let v_i be the number of available windows for document i, ordered by retrieval rank. We first give each document up to $F = 4$ windows, bounded by availability and the remaining budget. Any remaining budget is then distributed round-robin over documents in rank order until the cap is reached or no document has unused windows. If a document has more available windows than its quota, we take evenly spaced windows from that document.

Retain data. Following prior work [12, 18] we use `WikiText` as the retain set D_r, we keep this fixed across curators to isolate the effect of the forget set D_f. During training, each forget example is paired with a randomly sampled retain example when the unlearning objective uses retain regularization.

I Scaling Forget Request

See Table 6.

J Curator × Unlearner × Model Cross Evaluation

See Table 7. For each ⟨model, curator⟩ pair, the curated forget set D_f is constructed once and reused across SimNPO, UNDIAL, and RMU, so within each block only the unlearning algorithm changes. SimNPO rows repeat the corresponding entries of Table 1 (same runs), and the per-model cross evaluation for the EA curator is Table 2. Table 3 reports the three-model averages.

K Curator Output Sizes

The curators produce forget sets of different sizes; $|D_f|$ is a property of the curator under test, and we keep the unlearner's optimization hyperparameters (such as epochs) fixed irrespective of $|D_f|$.

Table 6: Forget-set size ablation at $|\mathcal{F}| = 100$, reported for Llama-3.1-8B and Qwen3-8B. Metrics and columns follow Table 1.

		Net reduction (%)		QA	Validation Δ (pp)						
	Method	F↑	R↓	Δ (pp)	GSM8K	BBH	WinoG	CoQA	HEval+	LMBDA	*Avg*
Llama-3.1-8B	EA (UNDIAL)	100.0%	93.0%	-4.0	+1.2	-0.6	-1.8	-4.5	+2.4	-4.0	-1.2
	EA (SimNPO)	100.0%	99.4%	-5.9	-2.8	-2.4	-3.4	+3.7	+3.0	-7.0	-1.5
	EA (RMU)	99.5%	94.0%	-1.7	-25.0	-58.0	-1.7	-48.7	+1.8	-45.6	-29.5
	Infinigram mid	49.0%	70.8%	-2.5	-0.9	-1.4	-1.6	+2.1	+1.8	-2.6	-0.4
	BM25 mid	38.8%	64.9%	-2.3	-1.9	-1.2	-1.7	+5.0	+4.3	-2.8	+0.3
Qwen3-8B	EA (UNDIAL)	100.0%	88.0%	-6.9	-2.7	-0.7	-0.1	-3.9	+0.6	-1.0	-1.3
	EA (SimNPO)	94.2%	72.8%	-10.9	-1.2	+2.0	+2.1	-4.7	-3.0	+1.1	-0.6
	EA (RMU)	45.8%	50.0%	-4.6	-2.6	-7.8	-1.7	-1.3	+4.9	-6.1	-2.4
	Infinigram mid	4.4%	3.4%	-6.6	-1.7	+1.5	+2.9	-1.6	-9.1	+3.4	-0.8
	BM25 mid	4.8%	3.9%	-6.7	-1.3	+1.4	+2.4	-2.2	-5.5	+3.3	-0.3

Table 8 reports the exact number of prefix–suffix windows produced by each curator at $|\mathcal{W}_f| = 50$. The retain set passed to the unlearner is fixed at 1,646 WikiText rows for every configuration.

L Songs vs. Books

Table 9 splits the Table 1 configurations by domain and averages over models. The qualitative result holds in both domains: retrieval curators do not produce strong selective suppression, and EA saturates target suppression without selectivity. Retain-side suppression under EA is higher for songs, which repeat internally and share choruses, idioms, and stock phrasing with other songs far more than books share text with other books. A 50-work request drawn from 4,616 songs and 50 books contains only one or two books on average, so the book column rests on few works per request.

Table 7: Per-model curator × unlearner cross evaluation at $|\mathcal{W}_f| = 50$. Metrics and columns follow Table 1.

			Net reduction (%)		QA	Validation Δ (pp)						
		Model	F↑	R↓	Δ (pp)	GSM8K	BBH	WinoG	CoQA	HEval+	LMBDA	*Avg*
BM25 pre	SimNPO	LLAMA-3.1-8B	28.9%	51.2%	-0.4	-0.1	+0.1	-0.6	+1.9	+1.2	-1.0	+0.3
		OLMO-3-7B	1.0%	2.9%	-0.1	+1.7	-0.1	-0.1	-3.8	-0.6	-0.0	-0.5
		QWEN3-8B	1.6%	2.1%	-3.9	-1.4	+1.4	+3.2	-3.5	-6.7	+3.8	-0.5
		Average	10.5%	18.7%	-1.5	+0.1	+0.5	+0.8	-1.8	-2.0	+0.9	-0.2
	UNDIAL	LLAMA-3.1-8B	100.0%	91.8%	-1.1	+0.8	-0.1	-0.6	-12.7	+0.6	-1.8	-2.3
		OLMO-3-7B	97.8%	79.6%	-0.1	+1.5	-0.1	+0.2	-5.1	+4.3	-1.3	-0.1
		QWEN3-8B	96.7%	79.0%	-3.5	-1.3	-5.1	+3.6	-5.3	-0.6	+0.3	-1.4
		Average	98.2%	83.5%	-1.6	+0.3	-1.8	+1.1	-7.7	+1.4	-0.9	-1.3
	RMU	LLAMA-3.1-8B	97.0%	94.7%	-7.4	-48.7	-64.6	-6.8	-48.1	-0.6	-62.2	-38.5
		OLMO-3-7B	-0.2%	6.9%	+0.3	+1.4	-0.0	+1.4	-0.5	+3.0	-1.3	+0.7
		QWEN3-8B	1.6%	5.0%	+0.1	-1.1	-0.4	-1.2	-0.6	+3.0	-0.8	-0.2
		Average	32.8%	35.5%	-2.3	-16.1	-21.7	-2.2	-16.4	+1.8	-21.4	-12.7
BM25 mid	SimNPO	LLAMA-3.1-8B	16.7%	54.0%	-1.1	-0.2	-0.4	-1.5	-2.4	+1.8	-1.4	-0.7
		OLMO-3-7B	1.5%	3.7%	-0.0	+1.0	-0.2	+0.0	-1.9	-1.2	-0.3	-0.4
		QWEN3-8B	2.3%	2.7%	-5.8	-1.1	+1.4	+3.6	-0.6	-10.4	+3.6	-0.6
		Average	6.8%	20.1%	-2.3	-0.1	+0.3	+0.7	-1.6	-3.3	+0.6	-0.6
	UNDIAL	LLAMA-3.1-8B	100.0%	99.0%	-2.5	-5.8	-4.2	-0.3	-18.4	-4.3	-3.3	-6.1
		OLMO-3-7B	86.9%	80.3%	-1.0	-1.0	-0.9	+0.5	-12.3	+2.4	-1.8	-2.2
		QWEN3-8B	91.9%	80.1%	-4.1	-2.2	-27.6	+2.8	-14.2	-2.4	-1.7	-7.5
		Average	92.9%	86.5%	-2.5	-3.0	-10.9	+1.0	-15.0	-1.4	-2.3	-5.3
	RMU	LLAMA-3.1-8B	94.3%	95.6%	-23.5	-48.8	-64.6	-17.4	-49.1	-24.4	-68.8	-45.5
		OLMO-3-7B	12.9%	39.8%	-6.4	-3.3	-26.7	-0.4	-20.7	-1.2	-35.4	-14.6
		QWEN3-8B	5.0%	9.3%	-4.0	-2.3	-5.1	-2.8	-0.1	+4.9	-2.9	-1.4
		Average	37.4%	48.2%	-11.3	-18.1	-32.1	-6.9	-23.3	-6.9	-35.7	-20.5
Infinigram mid	SimNPO	LLAMA-3.1-8B	19.4%	55.7%	-0.8	+0.3	-0.4	-0.9	+0.9	+0.0	-1.4	-0.2
		OLMO-3-7B	1.6%	3.2%	-0.1	+0.9	-0.3	+0.0	-0.8	-0.6	-0.1	-0.2
		QWEN3-8B	2.8%	2.8%	-4.6	-1.0	+1.5	+3.9	-3.1	-9.8	+3.6	-0.8
		Average	7.9%	20.6%	-1.8	+0.1	+0.3	+1.0	-1.0	-3.5	+0.7	-0.4
	UNDIAL	LLAMA-3.1-8B	99.7%	94.6%	-1.9	+1.0	-0.9	+0.2	-10.5	+3.0	-2.4	-1.6
		OLMO-3-7B	75.8%	69.4%	+0.1	+1.1	-0.8	+0.2	-8.3	+1.8	-1.4	-1.2
		QWEN3-8B	78.0%	72.9%	-4.5	-1.4	-10.5	+3.0	-6.7	-1.2	-2.4	-3.2
		Average	84.5%	79.0%	-2.1	+0.2	-4.1	+1.1	-8.5	+1.2	-2.1	-2.0
	RMU	LLAMA-3.1-8B	93.8%	92.5%	-16.6	-49.6	-64.6	-10.8	-49.1	-6.7	-65.3	-41.0
		OLMO-3-7B	6.5%	32.5%	-3.1	-1.1	-17.7	+0.0	-9.8	-0.6	-33.5	-10.5
		QWEN3-8B	4.0%	7.9%	-2.3	-2.1	-3.1	-2.7	-0.5	+3.0	-2.6	-1.3
		Average	34.8%	44.3%	-7.3	-17.6	-28.5	-4.5	-19.8	-1.4	-33.8	-17.6

Table 8: Number of curated forget-set windows at $|\mathcal{W}_f| = 50$.

Model	BM25 pre	BM25 mid	Infini-gram mid	EA
LLAMA-3.1-8B	2,901	6,350	5,787	8,513
OLMO-3-7B†	2,432	4,839	4,442	8,084
NEMOTRON-9B	2,835	6,324	6,011	8,840
QWEN3-8B	2,772	6,326	5,960	10,128
GEMMA-3-12B	3,028	6,323	6,096	27,936
OLMO-3-32B	2,737	6,381	6,187	67,323

†For OLMO-3-7B, only 38 works cleared the model-extractability threshold, so its forget pool contains 38 works rather than 50.

Table 9: Forget (F) and retain (R) net reduction (%) by domain, SimNPO, $|\mathcal{W}_f| = 50$, averaged over models.

Curator	Books		Songs	
	$F \uparrow$	$R \downarrow$	$F \uparrow$	$R \downarrow$
BM25 pre	8.0	25.4	23.0	29.9
BM25 mid	18.6	31.3	19.5	26.7
Infini-gram mid	26.1	30.8	20.7	27.4
EA	99.1	66.9	99.6	99.5

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: The abstract and Section 1 state our three contributions (formalizing forget-set curation, releasing CLEANSLATE, and identifying the two failure modes of natural and evaluation aware curators), and these correspond directly to the formal setup in Section 3, the benchmark in Section 5, and the experiments in Section 6.

   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: Section 7 (Discussion and Future Direction) states the scope of the study (verbatim output suppression for songs and books, not concept, entity, or factual unlearning), notes that our exact-overlap diagnostics miss paraphrase, translation, and semantic reuse, and lists future directions including dense or hybrid retrieval, influence estimation, and datamodeling-based curators.

   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: [N/A]

   Justification: The paper is empirical. Section 3 introduces notation for forget-set curation ($\mathcal{W}$, $\mathcal{W}_f$, $\mathcal{A}, \mathcal{U}$) but does not state or prove theorems.

   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: Section 5 specifies the extraction protocol (window length, stride, τ); Section 6 and Appendix I (and the algorithm/curation appendices) describe curators, unlearning objectives, hyperparameters, and the fixed WikiText retain set. Per-model extraction profiles, curated forget/retain sets, and the QA benchmark are released through the CLEANSLATE resources described in Section 5.

   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: [Yes]

   Justification: Code and data are linked from https://akhatua2.github.io/CleanSlate/.

   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: Section 6 and the appendix specify per-objective hyperparameters for GA, NPO, SimNPO, RMU, and UNDIAL, the FSDP training setup, forget-pool size $|\mathcal{F}|$, batching, the WikiText retain set, and the per-model extraction thresholds.

   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: Objects under evaluation is computed across mulitple models and averaged out.

   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: Please refer to Appendix Section B

   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 work conforms to the NeurIPS Code of Ethics. We use only publicly available pretrained models, do not release verbatim copyrighted text in our artifacts (we release content IDs and per-window extraction probabilities, not lyric/book text).

   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: Section 7 discusses positive impacts (auditing tools for copyright/privacy compliance, transparent unlearning evaluation) and negative impacts (extraction profiles could in principle aid targeted prompting; collateral damage in practical unlearning may produce false assurances of removal).

    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: [Yes]

Justification: CLEANSLATE releases content metadata, extraction probabilities, and curated indices rather than the full verbatim text of copyrighted songs/books. We do not release tools that increase extraction capability.

Guidelines:

- 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: CLEANSLATE includes copyrighted songs and some copyrighted books whose ownership is not with the authors. We acknowledge and credit this, one of the core motivations of our problem setting is copyright suppression. We also hold off on the release of the exact text of these works.

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: [Yes]

Justification: All new assets like the CLEANSLATE dataset construction process is provided.

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: The paper does not involve crowdsourcing or research with human subjects. QA pairs are generated with Gemini and filtered with Qwen and Llama judges, described in Section 5 and Appendix F.

    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: No human-subjects research was conducted.

    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: [Yes]

    Justification: LLMs are a core component of CleanSlate-QA construction: `gemini-3.1-flash-lite-preview` extracts propositions and generates QA pairs, and `Qwen/Qwen3.5-9B` and `meta-llama/Meta-Llama-3-8B-Instruct-Lite` filter and judge them (Section 5, Appendix F).

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.