SLEEPING SECRETS: HOW FINE-TUNING REAWAKENS PRIVACY RISKS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Beyond adapting Large Language Models (LLMs) to specialized applications, fine-tuning has recently been shown to recover private information that is no longer accessible through direct queries. Previous fine-tuning recovery attacks, however, require genuine private supervision drawn from the same distribution, i.e., the previous training dataset. We argue that such recovery remains possible without such impractical knowledge. We show that LLM-generated candidates can provide sufficient supervision to recover previously learned private associations. Based on this, we propose **ReGap**, a data-free attack that recovers private associations using task structure, filters them by answer-token likelihood, and updates the target model via low-rank adaptation. Specifically, **ReGap** requires neither target answers nor auxiliary genuine private supervision. Across six GPT-2, OPT, and Qwen3 models, **ReGap** improves target-association recovery by 6-21 percentage points over the post-training target model. Recovery remains substantial even when the adaptation identities are disjoint from all memorized and evaluation identities, with no exact target answers appearing in the generated or selected supervision. Moreover, the same trained adapters increase recovery from 42% to 63% on a previously exposed checkpoint, but produce no gain on a matched checkpoint that never encountered the targets. This contrast shows that adaptation alone is insufficient to explain the observed recovery and that prior target exposure strongly affects post-adaptation recoverability. Our findings highlight that routine model customization can reawaken latent privacy risks, warranting urgent attention from the academic and industrial communities.

1 INTRODUCTION

Fine-tuning has become a routine part of the lifecycle of Large Language Models (LLMs), enabling pretrained models to adapt to new tasks, domains, and applications (Gururangan et al., 2020; Hu et al., 2021). This continued adaptation also has implications for privacy. LLMs can memorize sensitive information and personally identifiable information (PII) from their pretraining data and disclose it in response to carefully crafted queries (Carlini et al., 2019; 2021; Tirumala et al., 2022; Biderman et al., 2023; Lukas et al., 2023). Recent work even shows that adaptation can introduce additional privacy risks: sensitive information may be unintentionally memorized during fine-tuning, model-generated training data can amplify privacy leakage, and even benign fine-tuning can weaken contextual privacy protections (Szep et al., 2026; Akkus et al., 2025; Goel et al., 2026a). These findings motivate evaluating privacy not only at a fixed checkpoint, but across the model's subsequent adaptation lifecycle.

A particularly relevant line of work studies whether later updates can restore access to information that is difficult to elicit directly. Janus shows that fine-tuning on genuine private association pairs can recover previously learned PII (Chen et al., 2024). Related studies show that downstream fine-tuning can also reverse apparent forgetting after machine unlearning (Wang et al., 2025), while malicious model-side interventions can increase privacy leakage during subsequent fine-tuning (Rashid et al., 2025). These results establish post-adaptation privacy as an important attack surface, but they leave a different attacker capability largely unexplored. Specifically, they assume access to authentic private records from the target training distribution Chen et al. (2024), an advantage that may be unavailable to an attacker who only obtains the released model. In this setting, the attacker has access to the released post-training model, but not to the original model from which it was derived.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This released model is our target model: the attacker seeks to recover private associations learned before post-training, without genuine private records as supervision.

This leads to our central question: can previously learned associations become more recoverable using only model-generated supervision, without externally supplied private records? We further ask whether recovery persists when adaptation uses neither target identities nor exact target answers. To investigate these questions, we propose **ReGap**, a recovery attack that generates candidate associations from the target model, selects candidates using answer-token likelihood under the same frozen model, and trains a fresh LoRA adapter on the resulting model-generated supervision. **ReGap** uses neither target answers nor auxiliary genuine private records during candidate generation, selection, or adaptation.

Across six GPT-2, OPT, and Qwen3 settings, one round of **ReGap** improves target-association recovery by 6-21 percentage points over the target model. Recovery remains substantial when adaptation identities are disjoint from all memorized and evaluation identities, with no exact target answer appearing in either the generated candidate pools or selected examples. Moreover, applying the same trained adapters to matched checkpoints yields a 21-point gain only when the model previously encountered the targets, while a target-unexposed checkpoint shows no gain. These controls indicate that the observed recovery depends on prior exposure rather than adaptation alone, while avoiding any claim that individual memories remain internally intact.

Our main contributions are summarized as follows:

- We identify a post-adaptation privacy risk in which previously learned private associations can become substantially more recoverable even without genuine private supervision. In the structured association setting studied here, recovery can persist when adaptation data are constructed without target identities and without exact target-answer replay, showing that direct behavioral inaccessibility alone is insufficient to characterize later privacy exposure.
- We introduce **ReGap**, a recovery attack that replaces genuine private supervision with model-generated candidates. **ReGap** follows a Generate–Select–Adapt pipeline: it generates task-consistent candidates from the target model, ranks them by answer-token likelihood under the frozen model, and trains a fresh LoRA adapter on the selected supervision. We further provide a local gradient-based analysis that characterizes when auxiliary updates can reduce target loss without identity or answer overlap and motivates the frozen-model selection rule.
- We systematically characterize the effect and its boundaries across GPT-2, OPT, and Qwen3 models. One round of **ReGap** improves target-association recovery by 6–21 percentage points across six model settings; matched controls further show substantial recovery with unseen adaptation identities and a strong dependence on prior target exposure. At the same time, recovery weakens sharply for opaque random bindings and after explicit NPO unlearning, establishing important limits on when model-generated supervision enables recovery.

2 Related Work

2.1 Memorization and Behavioral Accessibility

Language models can memorize specific training examples, with memorization affected by repetition, model scale, and training dynamics (Carlini et al., 2019; Tirumala et al., 2022; Carlini et al., 2022; Biderman et al., 2023). Memorization also extends beyond verbatim sequences to entity-attribute associations that are directly relevant to privacy (Zhou et al., 2024). However, whether previously learned information can be elicited from a checkpoint may change after subsequent updates (Jang et al., 2021; Jagielski et al., 2022).

Recent work further highlights the distinction between behavioral forgetting and information removal. Latent-memory analyses and unlearning audits find evidence that information can remain recoverable or decodable even after its direct accessibility has been substantially reduced (Duan et al., 2025; Goel et al., 2026b; Cohen et al., 2026). We therefore use *Hidden* purely as a behavioral notion: low direct recovery does not imply either deletion or intact internal storage.

2.2 Privacy Extraction from Language Models

Training-data extraction studies show that memorized text and PII can be elicited through prompting, prompt optimization, and large-scale querying (Carlini et al., 2021; Lukas et al., 2023; Ozdayi et al., 2023; Keum et al., 2025; Nasr et al., 2025). These attacks primarily operate through inference-time interaction with a fixed model checkpoint.

Importantly, successful prediction of a private attribute does not necessarily imply item-specific memorization, since lexical and distributional cues can also support reconstruction (Luo et al., 2026). This distinction is particularly relevant to our setting, where structured associations may contain exploitable regularities even when the corresponding target answers are unavailable to the attacker.

2.3 Post-Adaptation Privacy Risks

Privacy risk can change after a model undergoes subsequent adaptation. Post-training updates can substantially alter model behavior (Gururangan et al., 2020; Qi et al., 2024). This extends to privacy: fine-tuning on generated data can increase PII extraction and membership inference risk (Akkus et al., 2025), benign adaptation can degrade contextual privacy (Goel et al., 2026a), and changes involving one set of personal records can affect behavior on related information (Borkar et al., 2025).

Most closely related to our work, Janus moves beyond fixed-checkpoint extraction by using fine-tuning as the attack interface: genuine private association pairs are used to restore access to previously learned PII (Chen et al., 2024). A related line of work studies the interaction between forgetting and subsequent adaptation. Downstream fine-tuning can recover information targeted by machine unlearning, motivating unlearning procedures designed to remain invariant to later updates (Wang et al., 2025). Conversely, Rashid et al. show that a maliciously modified pretrained model can exploit unlearning to amplify privacy leakage during later victim fine-tuning (Rashid et al., 2025).

ReGap studies a different supervision regime. It neither designs an unlearning defense nor modifies an original model to implant a vulnerability. Instead, the attacker starts from a given target model in which previously learned associations have low direct recoverability and constructs supervision from the model itself, without access to target answers or genuine private association pairs. Our question is therefore not simply whether adaptation can alter privacy, but whether previously learned associations can become more accessible again when genuine private supervision is unavailable.

3 Problem Formulation

We focus on whether fine-tuning can make previously learned private associations accessible again after their direct recoverability has declined. Let $\mathcal{D}_{\mathrm{tar}} = \{(x_i, y_i)\}_{i=1}^{N}$, where $N \geq 1$, x_i is a query key, and y_i is a nonempty private attribute. A fixed task template τ gives $q_i = \tau(x_i)$ and $t_i = (q_i, y_i)$. Under a fixed evaluation protocol, let $\hat{y}_i(M)$ be the continuation generated by M for q_i, excluding the prompt. We measure recovery by

$$\mathrm{Rec}(M) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\mathrm{Norm}(y_i) \neq \varepsilon \ \wedge \ \mathrm{Norm}(y_i) \preceq_{\mathrm{sub}} \mathrm{Norm}(\hat{y}_i(M))] \,, \tag{1}$$

where ϵ is the empty string, $\preceq_{\mathrm{sub}}$ denotes contiguous substring inclusion, and Norm lowercases text, trims surrounding whitespace, and collapses consecutive whitespace. Prompting, decoding settings, and generation budgets are held fixed across models; any evaluation randomness is included in the evaluation protocol. Let M_{m} be an original model that learned the target associations. Subsequent non-target training produces a post-training model M_h with $\mathrm{Rec}(M_h) < \mathrm{Rec}(M_{\mathrm{m}})$ (Chen et al., 2024). This establishes neither deletion nor intact internal retention of the associations. The attacker receives only M_h.

Threat model. The attacker knows the association schema and answer format and can query the target model M_h, evaluate token-level likelihoods, and fine-tune it. Adaptation uses auxiliary keys X_{aux} and model-generated answers, without externally supplied target answers or genuine private association pairs. No distributional match between externally supplied auxiliary data and the targets is required. We use *data-free* in this supervision sense; generated candidates may nevertheless coincide with true answers. Correctness labels, target-recovery feedback, and M_{m} are unavailable for

attack construction or model selection. Ground truth is reserved for evaluation and post-hoc analysis. In the target-aware setting, X_{aux} may include known target keys. In the *Unseen-Identity Control*, it is disjoint from the keys used in the controlled memorization stage and target evaluation. Target identities are therefore unnecessary for constructing adaptation data, although evaluation queries still use the target keys.

Objective. Let $\mathcal{D}_{\text{aux}}$ denote the auxiliary fine-tuning dataset constructed by selecting model-generated QA pairs. With $\mathcal{K}_{\text{att}}$ denoting the task knowledge and auxiliary keys, attacker constructs

$$\mathcal{D}_{\text{aux}} = \mathcal{B}(M_h, \mathcal{K}_{\text{att}}), \quad M_a = \mathcal{A}(M_h, \mathcal{D}_{\text{aux}}). \tag{2}$$

Our objective is to determine whether the attacker can achieve

$$\Delta_{\text{rec}} = \text{Rec}(M_a) - \text{Rec}(M_h) > 0. \tag{3}$$

4 From Recovery Geometry to ReGap

ReGap constructs fine-tuning supervision from the target model itself through a Generate-Select-Adapt pipeline, as illustrated in Figure 1. We first characterize how an auxiliary update can improve unknown targets, then motivate an observable selection rule and specify the Generate-Select-Adapt procedure.

4.1 Local Geometry of Fine-Tuning

Let $M(\phi)$ be a parameterization of fine-tuning, with $\phi \in \mathbb{R}^d$ denoting displacements in the chosen trainable coordinates and $M(0) = M_h$. For $z = (q, y)$ with $|y| \geq 1$ answer tokens, define

$$\ell_z(\phi) = -\frac{1}{|y|} \sum_{t=1}^{|y|} \log p_{M(\phi)}(y_t \mid q, y_{<t}), \qquad g(z) = \nabla_\phi \ell_z(0). \tag{4}$$

Prompt tokens do not contribute to the loss. For a nonempty auxiliary set S, let

$$L_T(\phi) = \frac{1}{N} \sum_{i=1}^{N} \ell_{t_i}(\phi), \quad L_S(\phi) = \frac{1}{|S|} \sum_{z \in S} \ell_z(\phi), \tag{5}$$

and write $g_T = \nabla L_T(0)$ and $g_S = \nabla L_S(0)$. The local gradient-alignment kernel is

$$\kappa_h(z, z') = \langle g(z), g(z') \rangle. \tag{6}$$

Theorem 1 (Local target-loss change)**.** *If L_T is β-smooth on the segment from 0 to a parameter displacement δ, then*

$$L_T(\delta) \leq L_T(0) + \langle g_T, \delta \rangle + \frac{\beta}{2} \|\delta\|_2^2. \tag{7}$$

In particular, for the reference gradient step $\delta = -\eta g_S$, $\eta > 0$, smoothness on that segment gives

$$L_T(-\eta g_S) \leq L_T(0) - \frac{\eta}{N|S|} \sum_{i=1}^{N} \sum_{z \in S} \kappa_h(t_i, z) + \frac{\beta \eta^2}{2} \|g_S\|_2^2. \tag{8}$$

Thus, the target loss strictly decreases if

$$\langle g_T, g_S \rangle > \frac{\beta \eta}{2} \|g_S\|_2^2. \tag{9}$$

Theorem 1 separates answer overlap from update alignment: auxiliary examples need not reproduce target answers to decrease target loss. For a candidate distribution Q, the quantity $\sigma_i(Q) = \mathbb{E}_{z \sim Q}[\kappa_h(t_i, z)]$ measures first-order susceptibility to auxiliary gradient updates in the chosen coordinates. Appendix H.4 gives sufficient shared-gradient conditions for $\sigma_i(Q) > 0$ even when candidate identities and answers do not overlap the target. This motivates the unseen-identity control without requiring an intact internal representation of each record. The general bound applies to an actual parameter displacement, whereas the kernel expression specializes to a gradient step. Its trajectory-wise extension appears in Appendix H.3. A decrease in L_T alone does not imply an increase in Rec: Appendix H.8 supplies a decoding-specific sufficient condition and distinguishes new recoveries from lost recoveries.

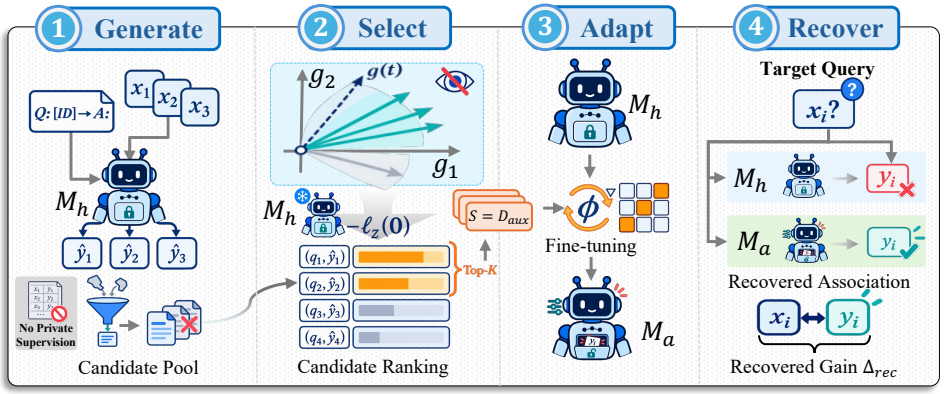

Figure 1: Overview of **ReGap**. Generate–Select–Adapt for recovering hidden associations.

4.2 From Self-Generated Answers to an Observable Selector

Why generation and selection matter. Self-generated supervision does not automatically produce a directed update. Under exact on-policy sampling and a matched complete-sequence likelihood, the expected gradient at the generating model is zero (Proposition 7). Answer-length weighting, biased decoding, and filtering need not preserve this cancellation. **ReGap** combines greedy and temperature-controlled top-p generation, format filtering, and support-based selection to shape the auxiliary data. These operations can create a nonzero mean update; whether that update benefits the targets depends on its alignment, not merely on its being nonzero.

An observable proxy for unavailable alignment. The first-order contribution of a candidate is

$$I_T(z) = \langle g_T, g(z) \rangle = \frac{1}{N} \sum_{i=1}^{N} \kappa_h(t_i, z). \tag{10}$$

This score requires unknown target answers. Instead, **ReGap** ranks candidates by their support under the unchanged target model,

$$s_h(z) = -\ell_z(0). \tag{11}$$

Support is observable but is neither a correctness label nor an alignment measurement. The following condition identifies when its ranking is useful.

Assumption 2 (Conditional support–alignment ordering)**.** For a random candidate pool $\mathcal{C} = (z_1, \ldots, z_m)$ of fixed size $m \geq K \geq 1$, let $\mathbf{s} = (s_h(z_1), \ldots, s_h(z_m))$ and $a_j(\mathbf{s}) = \mathbb{E}[I_T(z_j) \mid \mathbf{s}]$, with integrable $I_T(z_j)$. Almost surely, for every j, k,

$$s_h(z_j) \geq s_h(z_k) \quad \Longrightarrow \quad a_j(\mathbf{s}) \geq a_k(\mathbf{s}). \tag{12}$$

For variable-size pools, the condition is imposed conditional on pool size.

Proposition 3 (Conditional first-order selection)**.** *Under Assumption 2, selecting the K largest support scores maximizes $\mathbb{E}[\langle g_T, g_S \rangle \mid \mathbf{s}]$ among all selectors that choose exactly K candidates using only the scores and independent randomization.*

This is a conditional justification for the first-order selection criterion, not a guarantee that support ranks alignment in every model. Actual target-loss improvement additionally requires positive alignment large enough to dominate the update penalty in Theorem 1; Appendix H.7 makes this condition explicit. Comparison with selection from the same pool at random assesses the selector's practical value, while post-hoc gradient analysis can assess its alignment premise.

4.3 ReGap: Generate–Select–Adapt

Generate task-consistent candidates. For auxiliary keys $X_{\text{aux}} = \{u_a\}_{a=1}^{n_{\text{aux}}}$, construct $q_a^{\text{aux}} = \tau(u_a)$. Using a frozen copy of M_h, generate one greedy answer and J temperature-controlled top-p samples per query. Remove empty or malformed answers using only the known answer format, and

deduplicate identical normalized answers within each query. This yields a pool $\mathcal{C} = \{(q_a^{\text{aux}}, \hat{y}_{a,j})\}$ of m distinct query–answer candidates. No candidate is checked against a true private attribute. The same construction applies to target-aware and unseen-identity auxiliary keys.

Select by frozen support. For every $z = (q, \hat{y}) \in \mathcal{C}$, compute

$$\mathcal{L}_{\text{sel}}(z) = -\frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} \log p_{M_h}(\hat{y}_t \mid q, \hat{y}_{<t}) = \ell_z(0). \tag{13}$$

With a prescribed budget $K_0 \geq 1$, retain the $K = \min(K_0, m)$ candidates with smallest selection loss, breaking ties by a fixed rule. Selection is global across the pool and uses the target model likelihoods, not the temperature- or top-p-modified generation probabilities. If the filtered pool is empty, return M_h without adaptation. The selected set S constitutes the adaptation dataset, i.e., $\mathcal{D}_{\text{aux}} = S$.

Fine-tune from the target model. Starting from $\phi_0 = 0$, take H optimization steps on the selected answer-token objective,

$$L_S(\phi) = \frac{1}{K} \sum_{z \in S} \ell_z(\phi), \qquad M_a = M(\phi_H). \tag{14}$$

The loss averages tokens within each answer and then averages examples. Generation and scoring use the unchanged target model; only the model being adapted is updated. Generation, selection, and training budgets are fixed without target-answer validation or target-recovery feedback. This specifies one Generate–Select–Adapt round with multi-step fine-tuning, rather than an additional outer self-training loop.

5 Evaluation

We evaluate **ReGap** along three complementary questions: **(RQ1)** whether model-generated supervision enables recovery across models, **(RQ2)** how much target-specific information is required, and **(RQ3)** whether frozen-model selection contributes beyond random candidate selection. We additionally use a matched prior-exposure control to test whether the observed gains can be explained by ordinary relearning alone.

5.1 Experimental Setup

Benchmark. We primarily evaluate **ReGap** on 3,000 structured person-email associations derived from the Enron email corpus (Klimt & Yang, 2004) using the LLM-PBE pipeline (Li et al., 2024). We reserve 50 associations as genuine-private references and 100 as recovery targets. The remaining 2,850 associations form the non-target pool, of which 300 are used for non-target evaluation. We use WikiText-2 (Merity et al., 2016) for subsequent non-target continual training that produces the post-training model M_h. Enron person-email associations have been used in prior language-model privacy and recovery studies (Li et al., 2024; Chen et al., 2024), making them a natural testbed for our setting. We nevertheless treat evaluation on a single primary benchmark as a limitation and examine additional structural controls in the appendix.

Models. We evaluate six open-source autoregressive models spanning three families and approximately 0.1B-4B parameters: GPT-2 Small, Medium, and Large (Radford et al., 2019), OPT-1.3B (Zhang et al., 2022), and Qwen3-1.7B/4B (Yang et al., 2025). These models satisfy the access required by our threat model-token-level likelihood scoring and parameter-efficient adaptation-while providing both controlled within-family scaling through GPT-2 and cross-family evaluation through OPT and Qwen3.

Baselines. We use complementary baselines to isolate different sources of recovery. *Hidden* measures direct recoverability before adaptation. *Matched Random* samples the same number of examples from the same model-generated candidate pool and uses the same LoRA adaptation budget

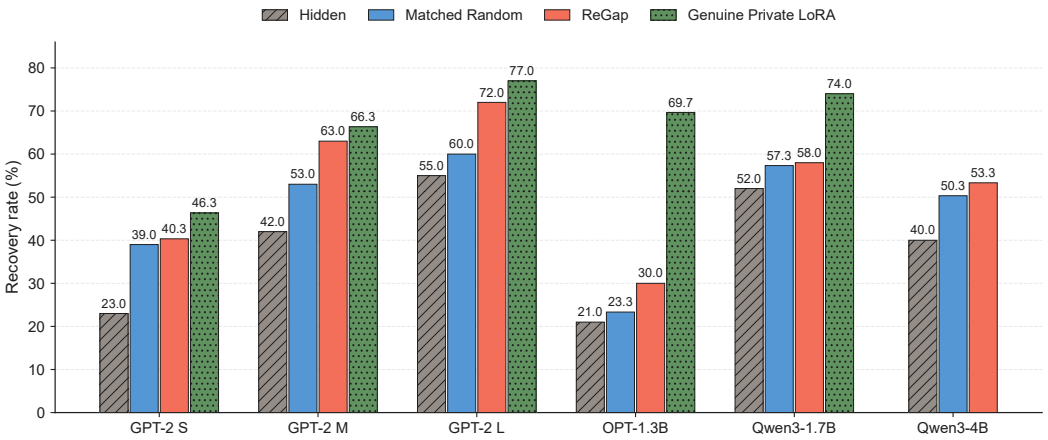

Figure 2: Target-association recovery across six model settings.

as ReGap, isolating the contribution of loss-based candidate selection. *Genuine-Private LoRA* replaces model-generated supervision with 50 genuine private-reference associations under the same adaptation budget, providing a stronger-supervision reference related to Janus (Chen et al., 2024). Matched Genuine-Private LoRA results are available for five model settings; for Qwen3-4B, a matched Genuine-Private LoRA run is not available, so we report an unmatched private-supervision result only as a descriptive reference. We additionally evaluate a parameter-free Janus-style in-context learning baseline on GPT-2 Medium and Large, with the full shot-wise results reported in the appendix. These baselines are chosen to match the post-adaptation recovery setting studied here. Inference-time extraction attacks instead operate on a fixed checkpoint without performing model adaptation, and therefore address a different attack interface rather than providing a directly matched comparison for **ReGap**. We use Hidden to represent this pre-adaptation accessibility, Genuine-Private LoRA to represent recovery with genuine private supervision, and Matched Random to isolate the effect of **ReGap**'s candidate-selection rule.

Metrics and Protocol. Our primary metric is target-association recovery, defined in Section 3 as the fraction of the 100 target answers that occur in the corresponding generated continuations after case normalization, whitespace trimming, and whitespace collapsing. Thus, one percentage point corresponds to one additional recovered target association. Unless otherwise stated, we use one Generate-Select-Adapt round with $K = 50$ selected candidates. Stochastic results report mean $\pm$ sample standard deviation over seeds 42, 43, and 44. Ground-truth target answers are used only for final evaluation and post-hoc auditing, never for candidate generation, scoring, selection, or adaptation.

5.2 RQ1: Recovery with Model-Generated Supervision

Figure 2 reports one-round recovery across all six model settings. **ReGap** improves over the unadapted target model in every setting, adding 6–21 recovered associations per 100 targets. The largest gain occurs on GPT-2 Medium, where recovery increases from 42/100 to 63/100; GPT-2 Large similarly increases from 55/100 to 72/100. The effect also appears on OPT and Qwen3, although its magnitude varies across models.

To quantify the cost of removing genuine private supervision, we additionally compare with a Genuine-Private Fine-Tuning setting, where the attacker is provided with the actual private associations. This setting provides genuine auxiliary private associations and serves as a stronger-supervision reference outside the ReGap threat model. While genuine private supervision achieves higher recovery overall, **ReGap** closes much of the gap on GPT-2 Medium and Large, with differences of 3.3 and 5.0 percentage points, respectively. These results show that model-generated supervision can recover substantial hidden associations without access to genuine private records. Full per-seed and ICL results are reported in the appendix.

5.3 RQ2: How Much Target-Specific Information Is Required?

We construct an Unseen-Identity Control whose adaptation identities are disjoint from all memorized and evaluation identities, while preserving the task schema and answer format. As shown in Table 1, **ReGap** improves recovery on both GPT-2 Medium and GPT-2 Large, increasing recovery by 15.3 and 14.0 percentage points over the target-model baseline, respectively. Across all evaluated seeds, the adaptation identities have zero overlap with the memorized and evaluation identities, and no exact target answers appear in either the generated candidate pools or selected examples ($0/1500$ candidate-pool hits and $0/150$ selected-example hits for each model). Thus, on GPT-2 Medium and Large, recovery improves without using target identities to construct the adaptation data and without exact target-answer replay in either the generated candidate pools or selected supervision. This rules out direct identity reuse and exact-answer copying as explanations for these gains, while leaving open the contribution of shared person–email regularities.

Table 1: Unseen-identity recovery.

Model	Hidden	**ReGap**	Gain
GPT-2 M	42.0	57.3 ± 1.5	+15.3
GPT-2 L	55.0	69.0 ± 1.0	+14.0

Table 1: Query ablation.

Condition	Recovery (%)
No adaptation	42.0
Correct query	63.0 ± 2.6
Shuffled query	63.0 ± 1.0
Generic query	41.0 ± 0.0

Query-information ablation. A matched ablation further clarifies what information is required for recovery. As shown in Table 5.3, shuffled entity-conditioned queries perform nearly identically to the original queries, whereas replacing them with a generic query reduces recovery to the pre-adaptation level. Thus, recovery does not require the exact target–query pairing, but benefits from preserving task- and entity-conditioned query information.

Prompt robustness. We further vary the surface form of the generation query under the same one-round Original-Loss protocol. As shown in Table 2, several English reformulations retain substantial recovery. The standard "What" prompt reaches $63.0 \pm 2.6\%$, while the "How" formulation and English paraphrase A both reach $61.7 \pm 1.2\%$. English paraphrase B is weaker at $54.7 \pm 2.3\%$, whereas the Chinese prompt remains near the pre-adaptation level at $41.7 \pm 0.6\%$. Thus, recovery is robust to several English reformulations, but is not invariant to prompt wording. Taken together, these controls narrow the information requirements of **ReGap**. Exact target identities and exact target–query pairings are not necessary for substantial recovery in the evaluated structured setting. However, removing most entity-conditioned query information eliminates the gain, and recovery remains sensitive to prompt formulation.

Table 2: Prompt robustness.

Prompt	Recovery (%)
What (standard)	63.0 ± 2.6
How	61.7 ± 1.2
Paraphrase A	61.7 ± 1.2
Paraphrase B	54.7 ± 2.3
Chinese	41.7 ± 0.6

5.4 RQ3: Does Candidate Selection Matter?

ReGap and Matched Random differ only in how adaptation examples are selected. As shown in Figure 2, **ReGap** achieves higher mean recovery in all six settings, with gains over Matched Random ranging from 0.7 to 12.0 percentage points. However, Matched Random itself improves over the pre-adaptation target model in every setting. Candidate generation and adaptation therefore account for a substantial part of the recovery effect, while frozen-model selection provides an additional, model-dependent amplification. We therefore view candidate selection as an amplifier rather than the sole source of recovery. Its benefit varies substantially across models, and we do not claim universal or statistically significant superiority over random selection.

5.5 Matched Prior-Exposure Control

To distinguish recovery from ordinary relearning, we apply the same trained adapters to a matched model checkpoint that never encountered the 100 target associations. No candidates are regenerated and no adapters are retrained, so the adaptation updates are fixed while prior target exposure changes.

As shown in Table 3, the same adapters increase recovery from 42.0% to 63.0% on the previously exposed model checkpoint, but produce no gain on the matched target-unexposed control, which remains at 13.0%. Matched Random shows the same qualitative dependence on prior exposure. Thus, adaptation alone is insufficient to explain the observed gain: post-adaptation recovery depends strongly on whether the model checkpoint had previously encountered the target associations.

Table 3: Matched prior-exposure control.

Checkpoint	Hidden	Random	**ReGap**
Prev. exposed	42.0	53.0	63.0
Target-unexp.	13.0	13.0	13.0

6 Discussion

Interpretation. Our results reveal a distinction between what a model exposes at a fixed checkpoint and what can become accessible after subsequent adaptation. The matched prior-exposure control is particularly informative: applying the same trained adapters increases recovery on a checkpoint that previously encountered the targets, but produces no gain on a matched checkpoint that never encountered them. Adaptation alone is therefore insufficient to explain the observed recovery. At the same time, this evidence is behavioral. Reduced direct recovery does not establish that the corresponding associations have been deleted, and successful post-adaptation recovery does not imply that individual memories remain internally intact or unchanged. **ReGap** instead shows that prior exposure can affect later recoverability under lightweight adaptation. The Unseen-Identity Control further shows that this effect need not depend on target-specific identities or exact target-answer replay, although knowledge of the underlying task structure remains available.

Boundaries, limitations, and defense implications. Our main evaluation focuses on structured Enron-derived person-email associations, whose identities and answers contain exploitable lexical and compositional regularities. Recovery weakens sharply when this structure is removed in the opaque random-binding control, indicating that model-generated supervision does not provide a generic mechanism for recovering arbitrary hidden associations. Recoverability also depends on how direct accessibility is reduced. Whereas our main target models are produced by non-target continual training, explicit NPO unlearning yields almost no additional recovery after **ReGap**: recovery changes from 71.0% to 71.3% under the moderate setting and remains at 32.0% under the stronger setting. This result identifies an important boundary rather than establishing general robustness to unlearning, but it also suggests that post-training mechanisms that directly suppress target-specific information may be more resistant to later recovery than ordinary non-target training. More broadly, **ReGap** assumes knowledge of the association task and expected answer format, and our experiments focus on open-source models that permit token-level scoring and parameter-efficient adaptation. Extending the analysis to less structured private information, additional unlearning and privacy-preserving defenses, and restricted fine-tuning interfaces remains future work.

7 Conclusion

We show that fine-tuning can restore access to previously learned private associations without genuine private supervision. Our attack, **ReGap**, constructs supervision from the target model's own generations, selects candidates using answer-token likelihood, and fine-tunes on the selected examples. Across GPT-2, OPT, and Qwen3 settings, a single round improves target-association recovery by 6-21 percentage points. Recovery also persists in controls with disjoint adaptation identities, where post-hoc audits find no exact target answers in either the candidate pools or selected supervision. These results demonstrate that recovery does not necessarily require reintroducing the target records during adaptation. Our controls further clarify the conditions under which this recovery occurs. A matched prior-exposure comparison supports a role for the model's training history, while recovery weakens sharply for opaque random bindings and after explicit NPO unlearning. The effectiveness of model-generated supervision thus depends on the associations being recovered and the updates the model has undergone. Nevertheless, the absence of externally supplied private records is insufficient to rule out recovery. Privacy evaluation should therefore assess both what a released model reveals directly and whether subsequent customization can make previously learned information accessible again, including when the customization data come from the model itself.

AI USE STATEMENT

Generative AI tools were used during this work to assist with research discussion, experimental design feedback, code development and debugging, result interpretation, translation, and manuscript drafting and editing. They were also used to help organize experimental records and improve the clarity and structure of the presentation.

All AI-assisted outputs were manually reviewed by the authors. Experimental results, numerical claims, methodological descriptions, and citations were checked against the corresponding code, logs, source files, and original references before inclusion in the manuscript. The authors take full responsibility for the final content of the paper.

REFERENCES

Atilla Akkus, Masoud Poorghaffar Aghdam, Mingjie Li, Junjie Chu, Michael Backes, Yuyang Zhang, and Sinem Sav. Generated data with fake privacy: Hidden dangers of fine-tuning large language models on generated data. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 8075–8093, 2025.

Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090, 2023.

Jaydeep Borkar, Matthew Jagielski, Katherine Lee, Niloofar Mireshghallah, David A Smith, and Christopher A Choquette-Choo. Privacy ripple effects from adding or removing personal information in language model training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18703–18726, 2025.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

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

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Xiaoyi Chen, Siyuan Tang, Rui Zhu, Shijun Yan, Lei Jin, Zihao Wang, Liya Su, Zhikun Zhang, XiaoFeng Wang, and Haixu Tang. The janus interface: How fine-tuning in large language models amplifies the privacy risks. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1285–1299, 2024.

Liran Cohen, Yaniv Nemcovsky, and Avi Mendelson. Remind: Memorization and unlearning in llms through the lens of input loss landscapes. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47955–47993, 2026.

Sunny Duan, Mikail Khona, Abhiram Iyer, Rylan Schaeffer, and Ila Fiete. Uncovering latent memories in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 27339–27361, 2025.

Anmol Goel, Cornelius Emde, Seong Joon Oh, Sangdoo Yun, and Martin Gubri. Privacy collapse: Benign fine-tuning can break contextual privacy in language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8870–8892, 2026a.

Anmol Goel, Alan Ritter, and Iryna Gurevych. Auditing language model unlearning via information decomposition. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 808–826, 2026b.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8342–8360, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, et al. Measuring forgetting of memorized training examples. *arXiv preprint arXiv:2207.00099*, 2022.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. *arXiv preprint arXiv:2110.03215*, 2021.

Seongho Keum, Dongwon Shin, Leo Marchyok, Sanghyun Hong, and Sooel Son. Private investigator: Extracting personally identifiable information from large language models using optimized prompts. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 8175–8194, 2025.

Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pp. 217–226. Springer, 2004.

Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, et al. Llm-pbe: Assessing data privacy in large language models. *arXiv preprint arXiv:2408.12787*, 2024.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE symposium on security and privacy (SP)*, pp. 346–363. IEEE, 2023.

Xiaoyu Luo, Yiyi Chen, Qiongxiu Li, and Johannes Bjerva. Do llms really memorize personally identifiable information? revisiting pii leakage with a cue-controlled memorization framework. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33841–33861, 2026.

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

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *International Conference on Learning Representations*, volume 2025, pp. 82363–82435, 2025.

Mustafa Ozdayi, Charith Peris, Jack FitzGerald, Christophe Dupuy, Jimit Majmudar, Haidar Khan, Rahil Parikh, and Rahul Gupta. Controlling the extraction of memorized data from large language models via prompt-tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1512–1521, 2023.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, volume 2024, pp. 30988–31043, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Md Rafi Ur Rashid, Jing Liu, Toshiaki Koike-Akino, Ye Wang, and Shagufta Mehnaz. Forget to flourish: Leveraging machine-unlearning on pretrained language models for privacy leakage. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20139–20147, 2025.

Marton Szep, Jorge Marin Ruiz, Georgios Kaissis, Paulina Seidl, Rüdiger von Eisenhart-Rothe, Florian Hinterwimmer, and Daniel Rueckert. Unintended memorization of sensitive information in fine-tuned language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6461–6480, 2026.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

Changsheng Wang, Yihua Zhang, Jinghan Jia, Parikshit Ram, Dennis Wei, Yuguang Yao, Soumyadeep Pal, Nathalie Baracaldo, and Sijia Liu. Invariance makes llm unlearning resilient even to unanticipated downstream fine-tuning. *arXiv preprint arXiv:2506.01339*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhenhong Zhou, Jiuyang Xiang, Chaomeng Chen, and Sen Su. Quantifying and analyzing entity-level memorization in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19741–19749, 2024.

APPENDIX

TABLE OF CONTENTS

A Benchmark and Data Construction 15

A.1 Enron-Derived Person-Email Benchmark 15
A.2 Data Splits and Experimental Roles 15
A.3 Unseen-Identity Control 15
A.4 Opaque Random-Binding Control 16
A.5 Additional Enron Controls 16

B Main Protocol and Implementation Details 17

B.1 Post-Training Model Construction 17
B.2 Candidate Generation 17
B.3 Original-Loss Candidate Selection 17
B.4 LoRA Adaptation 18
B.5 Model-Specific LoRA Target Modules 18
B.6 Random Baseline and Fairness Controls 18
B.7 Evaluation Protocol 19
B.8 Main-Protocol Summary 19

C Full Recovery Results 20

C.1 Per-Seed Results Across Six Models 20
C.2 Multi-Round Recovery Dynamics 20
C.3 LoRA vs. Full Fine-Tuning 21
C.4 Random-Seed Consistency 22
C.5 Genuine-Private and ICL Baselines 22

D Post-hoc Analysis 23

D.1 Candidate Compatibility and Gradient Alignment 23
D.2 Set-Level Gradient Comparison 23
D.3 Additional Selector Analysis 24

E Boundary Experiments 25

E.1 Opaque Random Bindings 25
E.2 Controlled Enron Associations 25
E.3 Non-Enron-Domain Associations 26
E.4 Explicit NPO Unlearning Stress Test 27

F Current-Protocol Boundary Controls 27

F.1 Opaque Random Binding (Current R1) 27
F.2 Controlled Enron (Current R1) 28

G Leakage and Reproducibility Audits 28

G.1 Identity-Overlap Audit 28

G.2 Exact Target-Answer Replay Audit 28

G.3 Frozen-Scorer and Fresh-Adapter Verification 29

G.4 Seed and Protocol Consistency 29

G.5 Result Provenance 30

H Theoretical Analysis and Proofs 30

H.1 Analytical Setup 30

H.2 Proof of Theorem 30

H.3 Actual Updates and Multi-Step Fine-Tuning 31

H.4 Shared Gradient Structure Without Identity Overlap 31

H.5 Self-Sampling and Answer-Length Weighting 31

H.6 Proof of Proposition 32

H.7 From Conditional Ranking to a Finite-Step Loss Bound 32

H.8 Connecting Loss Changes to Decoded Recovery 33

A BENCHMARK AND DATA CONSTRUCTION

This appendix provides additional details on the privacy benchmark and the controlled variants used throughout our experiments. Our main benchmark is constructed from real email addresses extracted from the Enron email corpus, while the corresponding person identifiers are deterministically derived from the email strings. We additionally construct several controlled variants to separate recovery driven by target-specific exposure from recovery enabled by the broader person-email structure.

A.1 ENRON-DERIVED PERSON-EMAIL BENCHMARK

We construct the main privacy benchmark from the Enron email corpus (Klimt & Yang, 2004), using the Enron data released through the LLM-PBE pipeline (Li et al., 2024). We concatenate the available textual fields for each record and extract email addresses using a deterministic regular-expression-based pipeline. Email addresses are lowercased and deduplicated before constructing structured associations.

For each extracted email address, we derive a corresponding person identifier from the local part of the address. Specifically, the local part is split on common separators such as periods, underscores, and hyphens, and up to three resulting tokens are converted into title-cased name components. For example, an address of the form `alice.smith@domain.com` is converted into an association resembling `Alice Smith` $\rightarrow$ `alice.smith@domain.com`.

This construction has two important properties. First, the email strings themselves originate from the underlying corpus rather than being randomly synthesized. Second, because the person identifiers are mechanically derived from the email local parts, the resulting benchmark contains lexical and compositional structure between the query identity and the target answer. This structure is intentional in the main benchmark and motivates the additional opaque-binding control described in Section A.4.

A.2 DATA SPLITS AND EXPERIMENTAL ROLES

The privacy training pool contains 3,000 structured person-email associations. We partition these associations into three roles: a small private-reference set, the target recovery set, and a larger set of additional non-target associations.

The private-reference set contains 50 associations and is used only in reference experiments that explicitly assume access to genuine private supervision. It is never used by **ReGap** for candidate generation, selection, or adaptation.

The target set contains 100 associations and defines the private information whose post-adaptation recoverability we evaluate. These target answers are withheld from the attacker and are used only for final evaluation and post-hoc auditing.

The remaining 2,850 associations form the additional non-target pool. A 300-example subset of this pool is used to measure retained performance on non-target person-email associations. This quantity should be interpreted as non-target association exact-match performance rather than as a general measure of downstream model utility.

To reduce direct accessibility of the target associations, we continue training the original model on non-private text from WikiText-2. The target associations are absent from this continual-training corpus. We refer to the resulting model as the post-training model. In our main experiments, its direct target-association recovery is substantially reduced relative to the original model. This reduction is behavioral and does not imply that the corresponding information has been deleted or remains internally unchanged.

A.3 UNSEEN-IDENTITY CONTROL

The primary attack setting assumes that the attacker knows the identities in the target queries while not knowing their private email addresses. To test whether these target-specific identities are necessary for recovery, we construct an *Unseen-Identity Control*.

Table 4: Summary of the main Enron-derived benchmark splits. Target answers are unavailable during the **ReGap** attack and are used only for evaluation and post-hoc auditing.

Split	Size	Role
Private-reference set	50	Genuine-private reference only
Target associations	100	Recovery evaluation
Additional non-target pool	2,850	Non-target training/evaluation pool
Non-target evaluation subset	300	Association-level retention evaluation
Total privacy training pool	3,000	Structured person-email associations

In this control, candidate generation and adaptation use independently generated identities that are disjoint from the identities appearing in the private-reference set, target set, non-target evaluation set, and the full memorized privacy pool. The person-email task schema and expected answer format remain unchanged. Thus, the control removes target identity overlap while preserving the broader structure of the task.

We explicitly audit identity and answer overlap after candidate generation. Across all evaluated seeds for GPT-2 Medium and GPT-2 Large, the unseen identities have zero overlap with memorized or evaluation identities. Moreover, neither the generated candidate pools nor the selected adaptation examples contain any exact target answer. These checks allow us to rule out exact target-answer replay in this control, while still acknowledging that the attacker retains knowledge of the person-email task structure.

A.4 Opaque Random-Binding Control

The main Enron-derived benchmark contains substantial lexical structure: person identifiers are derived from email local parts, making some name-answer regularities potentially exploitable even without access to the correct target answers. We therefore construct an opaque random-binding control to examine how **ReGap** behaves when this compositional structure is removed.

In this variant, identities are synthetically generated and independently paired with random 12-character local parts under the reserved `privacy.invalid` domain. The resulting associations preserve the abstract person-to-email input/output format while removing the lexical relationship between a person's name and the corresponding answer.

The opaque benchmark contains 3,000 associations, including 100 target associations constructed using the same split size as the main benchmark. Because the answers are independently assigned random strings, successful recovery cannot exploit the local-part structure present in the Enron-derived benchmark. We use this experiment as a structural boundary rather than as a strictly matched causal ablation, since the opaque branch uses its own independently trained checkpoints.

A.5 Additional Enron Controls

We also construct two additional Enron-derived variants for descriptive boundary analysis.

Controlled Enron. To reduce artifacts introduced by deriving identities directly from email local parts, we extract sender display names and sender email addresses from structured message metadata and retain clean one-to-one name-address associations. The construction scans 490,408 messages, yielding 1,184 raw candidate pairs and 757 clean one-to-one associations after filtering. This variant provides a setting in which the person identifier is obtained independently from the email local-part transformation used in the main benchmark.

Non-Enron-domain subset. We additionally remove addresses belonging to `enron.com` and its subdomains, obtaining 714 external-domain person-email associations. These examples still originate from messages in the Enron corpus and may retain lexical relationships between names and email addresses; therefore, we treat this experiment as a descriptive distributional control rather than as an unrelated-domain benchmark.

These additional variants are not used to define the headline results in the main paper. Instead, they help characterize how recovery changes as the structure and provenance of the person-email associations are modified.

B Main Protocol and Implementation Details

This appendix provides the implementation details omitted from the main paper, including construction of the post-training model, candidate generation, candidate scoring, LoRA adaptation, model-specific target modules, and the evaluation protocol.

B.1 Post-training Model Construction

Our experiments begin from a checkpoint that has learned the structured person-email associations described in Appendix A. We then perform non-target continual training on WikiText-2 (Merity et al., 2016), without including any of the target associations. The resulting model is the post-training model and serves as the target model for candidate generation, candidate scoring, and subsequent adaptation. In our main experiments, its direct target-association recovery is substantially reduced relative to the original model. This reduction is behavioral and does not imply that the target associations have been deleted from the model or remain internally unchanged.

For the prior-exposure control in Section 5.5, we additionally construct a matched target-unexposed checkpoint using the same overall training recipe while excluding the target associations from its entire training history. The same trained adapters are transferred to this checkpoint without additional candidate generation, candidate selection, or adapter training.

B.2 Candidate Generation

Candidate generation uses the frozen target model itself as the proposer. For each query, we generate one greedy completion together with five stochastically sampled completions. The decoding configuration is fixed across the main experiments and is summarized in Table 5.

Generated outputs are filtered for invalid responses, malformed answers, duplicate generations, and refusal-style outputs before candidate scoring. Importantly, this filtering depends only on output validity and format. A generated candidate is never accepted or rejected based on whether it matches the true target answer.

In the primary setting, generation queries contain the target identity but not its private answer. In the Unseen-Identity Control, the same generation procedure is applied to identities that are disjoint from all memorized and evaluation identities.

B.3 Original-Loss Candidate Selection

The main paper uses a single candidate-selection rule throughout: *answer-token average negative log-likelihood under the frozen target model*. For a prompt-candidate pair, prompt tokens are masked and only candidate answer tokens contribute to the score. Candidates with lower average negative log-likelihood are ranked more highly.

The same frozen target model is used both to generate and score candidates. No target-answer correctness labels, target gradients, or recovery-set performance are available to the selector. For the main one-round protocol, we retain the top

$$K = 50$$

lowest-loss candidates.

This selector corresponds to the *Original-Loss* configuration used in all headline results. Earlier experiments explored alternative selectors based on local loss-basin statistics and normalized relative gaps. Those historical variants are reported only where explicitly labeled in the appendix and are not used for the main six-model results.

Table 5: Candidate-generation configuration used in the main experiments. The target answers are never consulted when constructing or filtering the candidate pool.

Configuration	Value
Greedy generations per query	1
Sampled generations per query	5
Maximum new tokens	64
Temperature	0.7
Top-p	0.9
Repetition penalty	1.05
Target-answer verification	None

Table 6: LoRA adaptation configuration used for the main one-round experiments.

Hyperparameter	Value
LoRA rank	16
LoRA α	32
LoRA dropout	0.05
Learning rate	5×10^{-5}
Epochs	3
Batch size	1
Gradient accumulation	8
Gradient clipping	1.0
Selected examples	50
Training loss	Answer-token causal CE
Prompt tokens in loss	Masked

B.4 LoRA Adaptation

Each adaptation run starts from the same frozen target model and initializes a fresh LoRA adapter (Hu et al., 2021). Adapters are never continued across random seeds or reused across independent main experiments.

Training uses causal cross-entropy only on the generated answer tokens. Prompt tokens are masked from the adaptation loss. The main hyperparameters are shown in Table 6.

The main paper reports a single Generate-Select-Adapt round. Additional rounds are evaluated separately in Appendix C.2, where repeated adaptation does not consistently improve recovery across models.

B.5 Model-Specific LoRA Target Modules

We apply LoRA to the attention and feed-forward projections exposed by each model family. The target modules are listed in Table 7.

The evaluated model suite consists of GPT-2 Small, GPT-2 Medium, GPT-2 Large (Radford et al., 2019), OPT-1.3B (Zhang et al., 2022), Qwen3-1.7B, and Qwen3-4B (Yang et al., 2025).

B.6 Random Baseline and Fairness Controls

The matched Random Selection baseline uses candidates from the same generated candidate pool as **`ReGap`** and applies the same adaptation budget and LoRA configuration. The only difference is the candidate-selection rule: instead of ranking examples by frozen-model answer loss, the baseline samples the adaptation examples randomly from the corresponding candidate pool.

This design holds fixed the target model, generated candidate pool, number of selected examples, LoRA initialization procedure, number of epochs, and optimization hyperparameters. Consequently, the comparison isolates the contribution of frozen-model candidate selection from the broader effect of adapting on model-generated supervision.

Table 7: Model-specific LoRA target modules. All models use the same LoRA rank, scaling factor, dropout, and optimization configuration unless otherwise specified.

Model family	LoRA target modules
GPT-2	`c_attn`, `c_proj`, `c_fc`
OPT	`q_proj`, `k_proj`, `v_proj`, `out_proj`, `fc1`, `fc2`
Qwen3	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`

Table 8: Information available to the attacker in the main **ReGap** protocol. A checkmark indicates information used by the corresponding stage.

Information	Generate	Select	Adapt
Task structure	✓	✓	✓
Target identities	✓	✓	✓
Target model	✓	✓	✓
Generated candidates	-	✓	✓
Target answers	×	×	×
Real private auxiliary records	×	×	×
Target correctness labels	×	×	×
Target gradients	×	×	×

Each stochastic condition is evaluated with random seeds

$$42,\ 43,\ 44.$$

A fresh adapter is initialized independently for every seed.

B.7 Evaluation Protocol

Recovery is evaluated on the same 100 target associations used throughout the main experiments. For each target query, the input prompt is removed from the model output and the remaining full generated continuation is retained for evaluation. Evaluation uses greedy decoding with up to 48 newly generated tokens.

Both the generated continuation and the reference answer are normalized by lowercasing, trimming leading and trailing whitespace, and collapsing consecutive whitespace characters into a single space. A target is counted as recovered when the resulting nonempty normalized reference string occurs as a substring of the normalized generated continuation. Formally, the implemented decision rule is

$$\mathrm{hit}_i = \mathbf{1}[\mathrm{Norm}(y_i) \neq \epsilon \ \wedge\ \mathrm{Norm}(y_i) \text{ occurs in } \mathrm{Norm}(\hat{y}_i)]\,.$$

No email-specific regular expression is used to determine recovery. Email-format matching is computed separately and does not affect the primary recovery metric.

For stochastic adaptation results, we report the mean and sample standard deviation over seeds 42, 43, and 44. Ground-truth target answers are isolated from the attack pipeline and are used only for final recovery evaluation and post-hoc audits such as exact-answer replay analysis. They are not used for candidate generation, candidate scoring, candidate selection, or LoRA adaptation.

For the prior-exposure experiment, the previously exposed checkpoint and the matched target-unexposed checkpoint are evaluated on the same 100 target associations using the same decoding and recovery-evaluation procedure. The previously exposed checkpoint increases from 42.0% to 63.0% after **ReGap** adaptation, whereas the target-unexposed checkpoint remains at 13.0%.

B.8 Main-Protocol Summary

For clarity, Table 8 summarizes the information available at each stage of the main **ReGap** pipeline.

The Unseen-Identity Control further removes target identities from adaptation, while retaining the task structure and expected answer format.

Table 9: Per-seed one-round recovery under the current Original-Loss protocol. All values are recovery percentages on the same 100 target associations. Random Selection and **ReGap** use identical candidate pools, adaptation budgets, and LoRA configurations, differing only in the candidate-selection rule.

		Random Selection			ReGap			
Model	Hidden	Seed 42	Seed 43	Seed 44	Seed 42	Seed 43	Seed 44	**ReGap** Mean
GPT-2 Small	23.0	42.0	38.0	37.0	40.0	39.0	42.0	40.3
GPT-2 Medium	42.0	55.0	51.0	53.0	60.0	64.0	65.0	63.0
GPT-2 Large	55.0	63.0	59.0	58.0	72.0	71.0	73.0	72.0
OPT-1.3B	21.0	25.0	21.0	24.0	28.0	33.0	29.0	30.0
Qwen3-1.7B	52.0	57.0	59.0	56.0	59.0	59.0	56.0	58.0
Qwen3-4B	40.0	49.0	50.0	52.0	54.0	53.0	53.0	53.3

Table 10: Mean recovery across three adaptation rounds using the Original-Loss selector. Additional rounds are not uniformly beneficial across models.

Model	Round 1	Round 2	Round 3
GPT-2 Small	40.3	40.7	38.7
GPT-2 Medium	63.0	61.0	60.0
GPT-2 Large	72.0	69.7	68.0
OPT-1.3B	30.0	26.7	26.7
Qwen3-1.7B	58.0	58.7	57.7
Qwen3-4B	53.3	52.0	51.0

C Full Recovery Results

This section provides the complete recovery results underlying the aggregate numbers reported in the main paper. We first report per-seed one-round results for all six evaluated models, then examine multi-round dynamics, compare LoRA with full-parameter fine-tuning, and summarize additional random-seed diagnostics.

C.1 Per-Seed Results Across Six Models

Table 9 reports the individual recovery rates for each of the three random seeds used in the main one-round evaluation. The aggregate values exactly reproduce the means and sample standard deviations reported in the main paper.

The per-seed results show that recovery improvements are not driven by a single favorable run. **ReGap** exceeds the corresponding target model in every evaluated seed across all six model settings. Matched Random Selection also improves over Hidden in most settings, reinforcing the observation from the main paper that model-generated supervision itself contains useful recovery signal.

The additional benefit from loss-based selection remains model dependent. The mean **ReGap**-Random difference is 1.3 percentage points on GPT-2 Small, 10.0 on GPT-2 Medium, 12.0 on GPT-2 Large, 6.7 on OPT-1.3B, 0.7 on Qwen3-1.7B, and 3.0 on Qwen3-4B. We therefore interpret the selector as an additional filtering mechanism rather than as the sole source of recovery.

C.2 Multi-Round Recovery Dynamics

The main paper adopts a single Generate-Select-Adapt round. To examine whether repeated updates provide additional benefit, we also evaluate the Original-Loss procedure for up to three rounds. Table 10 reports the mean recovery at each round, and Figure 3 visualizes the corresponding multi-round dynamics.

Additional rounds do not consistently improve recovery. GPT-2 Medium and GPT-2 Large achieve their highest mean recovery after the first round and decline thereafter. OPT-1.3B similarly decreases after the first round. GPT-2 Small and Qwen3-1.7B show small second-round changes,

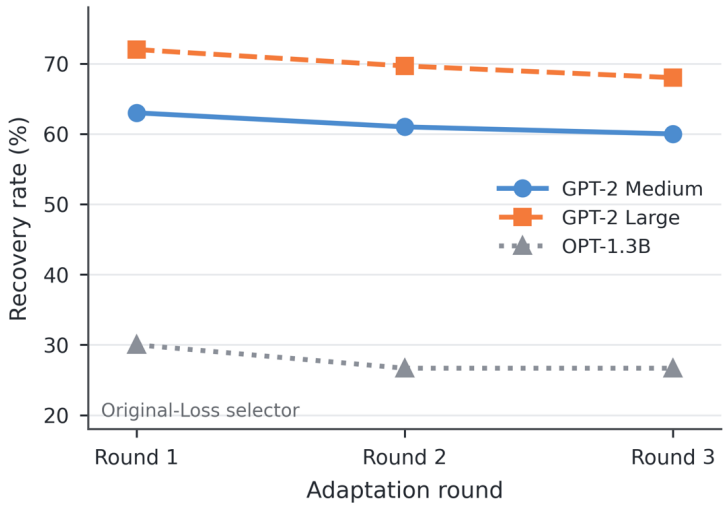

Figure 3: Multi-round recovery under the Original-Loss selector.

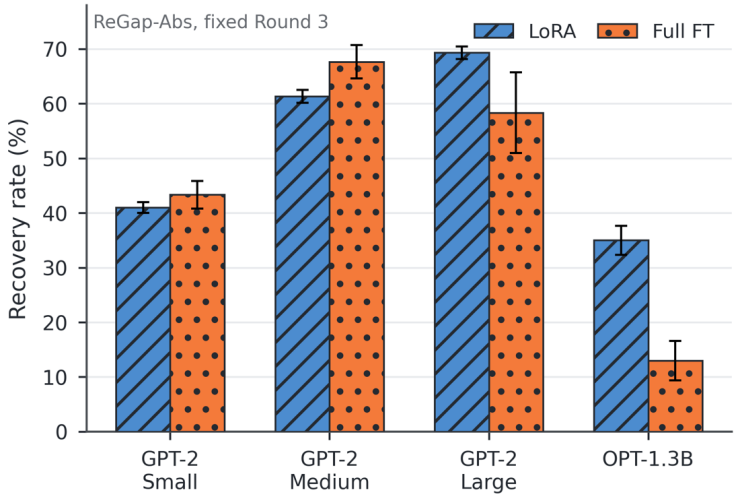

Figure 4: LoRA versus full fine-tuning under the historical ReGap-Abs protocol.

but neither exhibits sustained improvement by Round 3. These results motivate the one-round protocol used in the main paper: it captures most of the observable recovery while avoiding additional protocol complexity and non-monotonic updates.

C.3 LoRA vs. Full Fine-Tuning

We additionally examine whether recovery requires full-parameter optimization. This comparison was conducted in an earlier three-round **ReGap**-Abs protocol and is therefore reported as a diagnostic rather than as a direct extension of the current one-round Original-Loss results.

More trainable parameters do not uniformly increase recovery. As shown in Figure 4 and Table 11, full fine-tuning improves mean recovery for GPT-2 Small and Medium, but is substantially worse for GPT-2 Large and OPT-1.3B and exhibits larger variance in several settings. Although this historical experiment is not directly comparable to the current main protocol, it shows that lightweight LoRA adaptation is sufficient to produce substantial recovery and that increasing the number of trainable parameters does not guarantee stronger recovery.

Table 11: LoRA and full-parameter fine-tuning under the historical **ReGap**-Abs protocol. Values report mean ± sample standard deviation over three runs. This experiment uses a different selector and round configuration from the current main protocol and should be interpreted separately.

Model	LoRA	Full FT	Full FT − LoRA
GPT-2 Small	41.0±1.0	43.3±2.5	+2.3
GPT-2 Medium	61.3±1.2	67.7±3.1	+6.4
GPT-2 Large	69.3±1.2	58.3±7.4	−11.0
OPT-1.3B	35.0±2.7	13.0±3.6	−22.0

Table 12: Comparison with genuine-private supervision. Values are target recovery percentages; stochastic results report mean ± sample standard deviation over three seeds.

Model	Hidden	**ReGap**	Genuine-Private LoRA
GPT-2 Small	23.0	40.3 ± 1.5	46.3 ± 0.6
GPT-2 Medium	42.0	63.0 ± 2.6	66.3 ± 1.2
GPT-2 Large	55.0	72.0 ± 1.0	77.0 ± 1.0
OPT-1.3B	21.0	30.0 ± 2.7	69.7 ± 1.2
Qwen3-1.7B	52.0	58.0 ± 1.7	74.0 ± 1.0

C.4 Random-Seed Consistency

The headline Matched Random comparison uses the same three pre-specified seeds, 42, 43, and 44, across model settings. On GPT-2 Medium, the corresponding recovery rates are 55.0%, 51.0%, and 53.0%, yielding $53.0 \pm 2.0\%$. These are the values used in the main-paper comparison. Full per-seed results for all six models are reported in Table 9.

C.5 Genuine-Private and ICL Baselines

Genuine-Private LoRA. We compare **ReGap** with a stronger-supervision reference that replaces model-generated supervision with 50 genuine private-reference associations. For GPT-2 Small, Medium, and Large, OPT-1.3B, and Qwen3-1.7B, this baseline uses the same target model and LoRA adaptation budget as **ReGap**. It is therefore a Janus-style genuine-private reference rather than an exact reproduction of the original Janus protocol.

Genuine-private supervision is stronger overall, but the gap is model-dependent. On GPT-2 Medium and Large, Genuine-Private LoRA exceeds **ReGap** by only 3.3 and 5.0 percentage points, respectively, whereas the gap is substantially larger on OPT-1.3B and Qwen3-1.7B.

For Qwen3-4B, a matched Genuine-Private LoRA run is not available. The 66.0% private-supervision value shown in the main figure is included only as an unmatched descriptive reference and should not be interpreted as a matched result under the current protocol.

Janus-style in-context learning. We additionally evaluate whether genuine private examples can improve recovery without parameter updates. Following the comparison motivated by Janus (Chen et al., 2024), we use genuine private associations as in-context demonstrations and evaluate 1-, 5-, 10-, and 20-shot prompting on GPT-2 Medium and Large.

Under this matched prompting protocol, ICL does not improve recovery over the unadapted target model on either model. Increasing the number of demonstrations also does not monotonically improve recovery. We therefore use ICL as an additional parameter-free reference rather than as a substitute for the adaptation baselines.

Table 13: Janus-style in-context learning comparison. Values are recovery percentages; stochastic results report mean ± sample standard deviation over three seeds.

Method	GPT-2 Medium	GPT-2 Large
Hidden	42.0	55.0
ICL (1-shot)	22.0 ± 1.7	33.0 ± 2.0
ICL (5-shot)	6.7 ± 3.5	13.3 ± 11.5
ICL (10-shot)	10.0 ± 5.3	13.3 ± 5.5
ICL (20-shot)	10.7 ± 3.8	2.7 ± 1.2
ReGap	63.0 ± 2.6	72.0 ± 1.0
Genuine-Private LoRA	66.3 ± 1.2	77.0 ± 1.0

D Post-hoc Analysis

This section provides additional post-hoc analyses of the supervision selected during recovery. These experiments are intended to characterize empirical relationships between candidate compatibility, gradient direction, and recovery performance. They are diagnostic analyses rather than components of the **ReGap** attack itself: target gradients are never available during candidate generation, selection, or adaptation.

D.1 Candidate Compatibility and Gradient Alignment

We first examine whether candidates that appear more locally compatible with the target model also tend to induce update directions that are more aligned with target recovery.

For GPT-2 Medium, we analyze 100 candidate examples and compute two quantities post hoc. The first is a candidate-level local compatibility score derived from the frozen checkpoint. The second is the cosine similarity between the candidate-induced gradient and a target-recovery gradient computed only for analysis after the experiment has been completed.

As shown in Figure 5, candidate compatibility is positively associated with target-gradient alignment. We observe a Pearson correlation of

$$r = 0.747$$

and a Spearman rank correlation of

$$\rho = 0.716.$$

These results indicate that candidates scoring more favorably under the post-hoc compatibility measure also tend to induce update directions that are more aligned with the target-recovery direction.

Interpretation. This correlation provides evidence that local model compatibility is related to the usefulness of candidate supervision for recovery. However, it does not establish a causal mechanism, and the target gradient used in this analysis is not available to the attack. The main **ReGap** selector remains the answer-token negative log-likelihood computed solely from the frozen target model.

D.2 Set-Level Gradient Comparison

We next compare the aggregate gradient direction induced by several candidate sets. For each set, we compute the cosine similarity between the update direction produced by the selected supervision and the target-recovery gradient.

This analysis was conducted using an earlier selector-analysis protocol and is reported only as a post-hoc diagnostic.

As shown in Table 14, the aggregate direction induced by randomly selected candidates is nearly orthogonal to the target-recovery gradient, while the historical compatibility-based selection produces substantially stronger positive alignment. The genuine-private reference provides an upper reference point in this diagnostic.

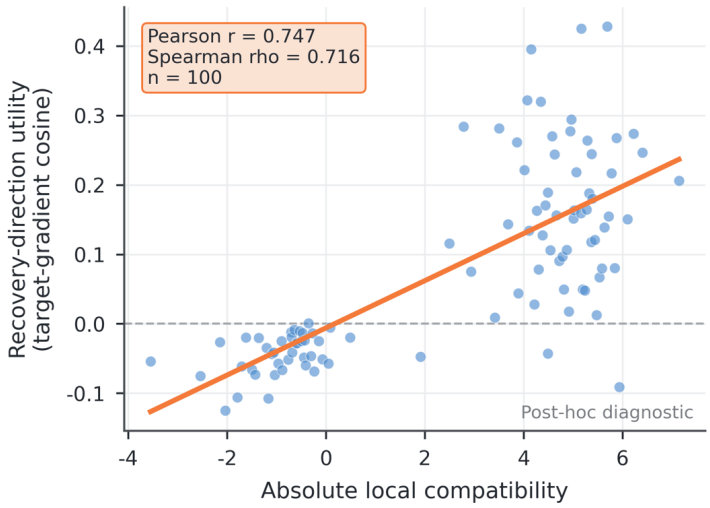

Figure 5: Candidate compatibility versus target-gradient alignment on GPT-2 Medium.

Table 14: Post-hoc set-level gradient cosine similarity on GPT-2 Medium. Higher values indicate stronger alignment with the target-recovery direction. The genuine-private set is shown only as a reference and is unavailable under the **ReGap** threat model.

Supervision set	Gradient cosine
Bottom-ranked candidates	−0.077
Random candidates	0.029
Historical **ReGap**-Abs selection	0.633
Genuine-private reference	0.805

These values are consistent with the hypothesis that candidate selection can concentrate supervision whose induced updates are more compatible with target recovery. We emphasize that this comparison is post hoc and does not imply that **ReGap** estimates or optimizes target gradients during the attack.

D.3 Additional Selector Analysis

Before adopting the simple Original-Loss selector used in the main paper, we evaluated several alternative candidate-ranking strategies. These experiments belong to an earlier selector-development protocol and are included to document the robustness of the broader candidate-selection idea rather than to redefine the current method.

The evaluated alternatives include a relative local-loss-gap score, a local-basin score, and an automatically searched selector configuration. Their best observed recovery rates on GPT-2 Medium are summarized in Table 15.

The more elaborate selectors do not provide a consistent advantage over the best manually specified alternatives. In particular, automatic selector search does not reliably improve upon simpler compatibility-based scores.

These observations motivated a simpler presentation in the final paper: the headline method uses frozen-model answer-token loss, which is easy to define, inexpensive to compute, and does not require perturbation-based statistics or target-side supervision. The historical selector experiments are retained here only as evidence that the recovery phenomenon is not tied to a single highly tuned ranking heuristic.

Table 15: Historical selector-development results on GPT-2 Medium. These values come from an earlier protocol and are not directly comparable to the current one-round Original-Loss headline results.

Selector variant	Best recovery (%)
Relative local-gap	64.7
Local-basin	64.4
Automatic selector search	64.0

E BOUNDARY EXPERIMENTS

The main experiments establish that model-generated supervision can increase post-adaptation recovery in structured person-email tasks. We next examine where this effect weakens or disappears. These experiments vary the amount and source of exploitable structure in the underlying associations.

Unless otherwise noted, the experiments in this section were conducted under earlier **ReGap**-Abs protocols and use independently trained checkpoints. They are therefore intended as descriptive boundary analyses rather than strictly matched extensions of the current one-round Original-Loss protocol.

E.1 OPAQUE RANDOM BINDINGS

A potential alternative explanation for recovery on person-email tasks is that the model can exploit lexical or compositional regularities between a person's identity and the corresponding email address. To remove these cues, we construct a non-compositional benchmark in which synthetically generated identities are independently paired with random 12-character local parts under the reserved `privacy.invalid` domain.

The abstract task format remains person-to-email, but the target answer has no designed lexical relationship with the entity name. As summarized in Tables 16 and 17, the previously exposed opaque condition drops from 95.0% recovery in the original model to 1.0% after non-target continual training and reaches only 2.0% after ReGap-Abs, while the matched target-unexposed branch remains at 0%.

No exact target answer appears in either the generated candidate pools or the selected pseudo-supervision in this experiment. Thus, the weak recovery cannot be attributed to direct replay of correct target answers.

Opaque bindings define an empirical boundary. Previously memorized arbitrary bindings show almost no answer-free recovery after direct accessibility is reduced. This result contrasts with the much larger gains observed in structured person-email settings and is consistent with exploitable association structure playing an important role in recovery. It does not establish that structure is the only factor determining recoverability, since the opaque branch uses its own independently trained checkpoint and historical adaptation protocol.

E.2 CONTROLLED ENRON ASSOCIATIONS

The main benchmark derives person identifiers mechanically from email local-parts, creating an explicit lexical relationship between query identities and target answers. To examine recovery when names and addresses originate from separate fields, we additionally construct a Controlled Enron benchmark from structured message metadata.

Sender display names or `X-From` fields are paired with the corresponding sender email addresses from the same message. After filtering, the dataset contains 757 clean one-to-one person-email associations.

Under the historical three-round **ReGap**-Abs protocol, recovery increases from 27.0% at the target model to 77.7$\pm$1.5%. Random model-generated supervision also produces substantial recovery at

Table 16: Boundary comparison between structured associations and opaque random bindings under the historical three-round **ReGap**-Abs protocol.

Association setting	Recovery rate (%)
Structured associations	61.3
Opaque random bindings	2.0

Table 17: Recovery on opaque non-compositional bindings. The experiment uses the historical three-round **ReGap**-Abs protocol. Values are percentages.

Exposure condition	Memorized	Hidden	**ReGap**
Previously Exposed	95.0	1.0	2.0
Target-Unexposed	-	0.0	0.0

66.7±2.3%, while Original-Loss selection reaches 75.7±2.1%. A genuine-private reference reaches 98.0%.

The Controlled Enron experiment suggests that the observed recovery is not restricted to associations constructed by mechanically converting email local-parts into person identifiers. However, it does not establish that all real-world person-email distributions support similarly strong recovery.

Recovery is not explained solely by direct candidate replay. We further examine targets whose correct answers are absent from their own first-round candidate pools. On this subset, recovery increases from 0.7% at the target model to 66.0% after the third **ReGap** round. Across all three seeds, none of these target answers appears in another target's first-round candidate pool or selected pseudo-supervision.

This audit rules out direct first-round exact-answer replay for this subset. It does not, however, establish a particular internal memory mechanism or rule out the use of broader distributional regularities.

E.3 Non-Enron-Domain Associations

The Controlled Enron data contain a strong corporate-domain regularity: many target addresses belong to `enron.com`. We therefore construct an additional subset that excludes addresses from `enron.com` and its subdomains. The resulting dataset contains 714 external-domain person-email associations.

Removing the dominant corporate-domain regularity sharply attenuates the answer-free recovery effect. As shown in Table 18, the target model achieves 46.0% recovery, while **ReGap**-Abs reaches 46.7±3.2%, corresponding to only a small mean increase. Matched Random Selection reaches 50.3±2.1%, and Original-Loss selection reaches 42.0±9.9%.

In contrast, the genuine-private reference reaches 80.7±1.2%, showing that substantial adaptation headroom remains even when answer-free model-generated supervision provides little additional recovery.

The benefit of model-generated supervision is distribution dependent. The contrast between Controlled Enron and Non-Enron suggests that the strength of answer-free recovery depends substantially on the statistical structure available in the association distribution. Removing the dominant domain pattern weakens the effect despite retaining the same broad person-email task format.

We interpret this result conservatively. The Controlled Enron and Non-Enron checkpoints are constructed independently, and the comparison is therefore not a strictly matched intervention on a single trained model. Nevertheless, together with the opaque-binding experiment, it provides descriptive evidence that **ReGap** does not act as a generic mechanism for recovering arbitrary hidden associations.

Table 18: Controlled Enron and Non-Enron recovery under the historical three-round protocol. The two settings use independently constructed checkpoints and are reported as descriptive boundary evidence rather than as a matched causal ablation. Values are percentages and report mean ± sample standard deviation where applicable.

Setting	Hidden	Random	Original Loss	**ReGap**-Abs	Genuine-Private
Controlled Enron	27.0	66.7±2.3	75.7±2.1	77.7±1.5	98.0±0.0
Non-Enron	46.0	50.3±2.1	42.0±9.9	46.7±3.2	80.7±1.2

Table 19: Recovery after explicit NPO unlearning on GPT-2 Medium. Values are percentages; stochastic results report mean ± sample standard deviation over three seeds.

NPO setting	Before adaptation	Matched Random	**ReGap**
Moderate	71.0	72.0 ± 0.0	71.3 ± 0.6
Strong	32.0	32.0 ± 0.0	32.0 ± 0.0

E.4 Explicit NPO Unlearning Stress Test

Our main target models reduce direct target accessibility through non-target continual training. We additionally consider a distinct setting in which GPT-2 Medium is first modified by Negative Preference Optimization (NPO), producing two defended checkpoints with different levels of direct target recoverability, denoted Moderate and Strong.

Unlike the historical experiments in Sections E.1–E.3, the recovery stage of this stress test uses the current one-round Original-Loss R1 protocol. Starting independently from each NPO checkpoint, we compare the unadapted model with Matched Random and **ReGap** under the same one-round adaptation procedure used in the main experiments. The NPO checkpoints themselves are produced by an explicit defense objective rather than by the non-target continual-training procedure used to construct the main target models. We therefore treat this experiment as a boundary stress test under a different accessibility-reduction mechanism, rather than as a strictly matched causal comparison with the main setting.

As shown in Table 19, **ReGap** provides almost no additional recovery from either NPO checkpoint. Under the Moderate setting, recovery changes from 71.0% before adaptation to $71.3 \pm 0.6\%$ after **ReGap**. Under the Strong setting, recovery remains unchanged at 32.0%. Matched Random exhibits the same qualitative behavior.

These results show that the substantial recovery observed after ordinary non-target continual training does not automatically extend to checkpoints produced by explicit NPO. Because the NPO checkpoints arise from a different accessibility-reduction mechanism, this experiment should be interpreted as a boundary stress test rather than as a matched comparison between continual training and NPO. It also covers only one model and one unlearning method and therefore does not establish general robustness to unlearning defenses.

F Current-Protocol Boundary Controls

To further characterize the boundary conditions of **ReGap** under the current one-round Original-Loss protocol, we evaluate two additional controls. Unlike the historical boundary experiments in Appendix E, these experiments follow the same evaluation protocol as the main study: one Generate-Select-Adapt round, Original-Loss candidate selection, and three random seeds.

F.1 Opaque Random Binding under Current R1

We evaluate **ReGap** on an Opaque random-binding benchmark where the association structure is preserved but semantic information about the target mapping is removed. This setting tests whether recovery can occur when the model itself cannot generate informative target-specific supervision.

Table 20: Current-protocol Opaque random-binding control.

Setting	Hidden	Random	**ReGap**
GPT-2 Medium	1.0	1.7±0.6	1.0±0.0

Table 21: Current-protocol Controlled Enron boundary control.

Setting	Hidden	Random	**ReGap**
GPT-2 Large	29.0	36.0±1.7	84.0±2.6

Across all three seeds, no exact target answer appears in either the generated candidate pools or the selected supervision examples. As shown in Table 20, **ReGap** does not improve over the target model in this setting. This result suggests that adaptation alone is insufficient: recovery requires model-generated supervision signals that preserve target-specific information.

F.2 Controlled Enron under Current R1

We additionally evaluate a Controlled Enron association benchmark under the current protocol. This experiment uses a GPT-2 Large post-hoc target model; therefore, it is treated as a supplementary boundary analysis rather than a strict matched-scale comparison with the main experiments.

As shown in Table 21, **ReGap** substantially increases recovery in this setting. However, the strong recovery is accompanied by target-answer replay in the generated supervision. Across seeds, each candidate pool contains 49-51 exact target answers, and the selected supervision contains 25-27 target answers. Therefore, this result should be interpreted as recovery enabled by model-generated target supervision rather than fully answer-free recovery.

G Leakage and Reproducibility Audits

This section summarizes the audits used to verify that the reported recovery results are not artifacts of target leakage, accidental answer replay, checkpoint mismatch, or seed reuse. We distinguish audits of the main target-identity-aware setting from those of the Unseen-Identity Control, because exact target-answer replay is ruled out only in the latter.

G.1 Identity-Overlap Audit

For the Unseen-Identity Control, adaptation identities are generated independently from the privacy benchmark. We explicitly compare them against all identity sets used in the experiment: the private-reference set, target set, non-target evaluation set, and the complete memorized privacy pool.

Thus, the recovery observed in the Unseen-Identity Control cannot be attributed to reusing target or memorized identities during adaptation. The control nevertheless preserves the person-email task structure and expected answer format, and should therefore not be interpreted as a task-unaware setting.

G.2 Exact Target-Answer Replay Audit

We next audit whether the generated or selected adaptation examples contain exact target answers.

For the Unseen-Identity Control, no exact target answer is found in either the generated candidate pools or the selected adaptation examples across all three evaluated seeds for GPT-2 Medium and GPT-2 Large. Table 22 summarizes these overlap and replay checks, all of which are zero across the evaluated seeds.

This result rules out exact target-answer replay as an explanation for the Unseen-Identity recovery effect.

Table 22: Leakage audit for the Unseen-Identity Control. All overlap and replay counts are zero across evaluated seeds.

Audit criterion	Result
Private-reference identity overlap	0
Target identity overlap	0
Non-target evaluation identity overlap	0
All memorized privacy identity overlap	0
Candidate-pool target-answer hits	0
Selected-example target-answer hits	0

Importantly, this conclusion does not extend to the primary target-identity-aware setting. In that setting, generated candidates can occasionally match target answers exactly, and selection can concentrate some of these candidates. We therefore do not claim that exact replay is absent from the main target-aware attack. Instead, the Unseen-Identity Control provides the cleaner evidence that substantial recovery can occur even when exact target answers never appear in the generated or selected supervision.

G.3 Frozen-Scorer and Fresh-Adapter Verification

The main **ReGap** protocol separates candidate generation and scoring from adaptation.

For every main experiment:

- candidate generation uses the fixed target model;
- candidate scoring uses the same frozen target model;
- target answers are not used during generation or scoring;
- each adaptation run initializes a fresh LoRA adapter from the same target model;
- adapters are not continued across random seeds; and
- the final adapted model is evaluated only after training is complete.

The matched Random Selection baseline follows the same protocol except for the candidate-selection rule. It uses the same generated candidate pool, the same number of selected examples, the same target model, and the same LoRA optimization configuration.

These checks ensure that the **ReGap**-Random comparison primarily changes which generated examples are selected for adaptation, while holding the training budget and model checkpoint fixed.

G.4 Seed and Protocol Consistency

All headline stochastic results use seeds 42, 43, and 44 under the current one-round Original-Loss protocol. As shown in Table 9, for GPT-2 Medium, the matched Random recoveries are 55.0%, 51.0%, and 53.0%, giving $53.0 \pm 2.0\%$. The corresponding **ReGap** recoveries are 60.0%, 64.0%, and 65.0%, giving $63.0 \pm 2.6\%$.

We additionally verify the current prior-exposure experiment using the same 100 target associations and the same one-round Original-Loss protocol. As shown in Table 23, on the previously exposed target model, **ReGap** increases recovery from 42.0% to $63.0 \pm 2.6\%$, while Matched Random reaches $53.0 \pm 2.0\%$. On the target-unexposed control, both remain at 13.0% for all three seeds.

The reproduced aggregate results exactly match the corresponding main-paper numbers. The current prior-exposure control therefore uses the same Original-Loss R1 adapters as the main GPT-2 Medium experiment rather than the historical **ReGap**-Norm adapters used in earlier development runs.

Table 23: Reproduction audit for the current prior-exposure control. All conditions use the same 100 target associations and the same evaluation protocol. Values are percentages.

Model checkpoint	Baseline	**ReGap** R1	Random R1
Previously Exposed	42.0	63.0±2.6	53.0±2.0
Target-Unexposed Control	13.0	13.0±0.0	13.0±0.0

G.5 Result Provenance

All headline results were verified against the corresponding audited current-protocol experiment outputs before inclusion in the paper. The six-model main results use the current one-round Original-Loss protocol, with the evaluation procedure and random seeds specified in Appendix B.

Historical experiments using earlier selectors, multi-round configurations, or independently constructed checkpoints are reported only where explicitly labeled as historical or diagnostic. These results are kept separate from the current-protocol headline results and are not used to replace or augment the primary comparisons.

Together with the identity-overlap, exact-answer replay, frozen-scorer, fresh-adapter, seed, and checkpoint audits above, this protocol separation provides additional checks against accidental leakage, mismatched adaptation budgets, adapter reuse across seeds, and mixing results obtained under different experimental configurations.

H Theoretical Analysis and Proofs

H.1 Analytical Setup

All gradients are taken in the fixed trainable coordinates of $M(\phi)$. Candidate texts, prompts, tokenizations, and selection decisions are held fixed when differentiating a training loss. In particular, we do not differentiate through generation or selection. The target examples are fixed analytical objects and are unavailable to the attack.

For a differentiable function F, β-smoothness on a segment means $\|\nabla F(u) - \nabla F(v)\|_2 \le \beta\|u - v\|_2$ for points u, v on that segment. Results involving expectations assume the stated first or second moments exist. Bounds on random updates additionally require a common smoothness constant on the corresponding update segments almost surely. No convexity assumption is used.

H.2 Proof of Theorem 1

Proof. By the fundamental theorem of calculus,

$$L_T(\delta) - L_T(0) = \int_0^1 \langle \nabla L_T(t\delta), \delta \rangle \, dt \tag{15}$$

$$= \langle g_T, \delta \rangle + \int_0^1 \langle \nabla L_T(t\delta) - g_T, \delta \rangle \, dt. \tag{16}$$

Cauchy–Schwarz and smoothness bound the integrand in the second term by $\beta t\|\delta\|_2^2$. Integration gives Equation 7. Substituting $\delta = -\eta g_S$ and using

$$\langle g_T, g_S \rangle = \left\langle \frac{1}{N}\sum_{i=1}^{N} g(t_i), \frac{1}{|S|}\sum_{z\in S} g(z) \right\rangle = \frac{1}{N|S|}\sum_{i=1}^{N}\sum_{z\in S} \kappa_h(t_i, z) \tag{17}$$

gives Equation 8. Since $\eta > 0$, Equation 9 makes its right-hand side strictly smaller than $L_T(0)$. □

H.3 Actual Updates and Multi-Step Fine-Tuning

Corollary 4 (Trajectory-wise target-loss bound)**.** *Let $\phi_{h+1} = \phi_h + \delta_h$ for $h = 0, \dots, H-1$. If L_T is β_h-smooth on each update segment, then*

$$L_T(\phi_H) - L_T(\phi_0) \le \sum_{h=0}^{H-1} \left(\langle \nabla L_T(\phi_h), \delta_h \rangle + \frac{\beta_h}{2} \|\delta_h\|_2^2 \right). \tag{18}$$

Proof. Apply the integral argument above at ϕ_h with displacement δ_h and sum the resulting inequalities. The left-hand sides telescope. □

This statement uses the actual optimizer displacements, including any momentum, preconditioning, or weight decay. For the preconditioned update $\delta_h = -\eta_h P_h \nabla L_S(\phi_h)$, the first-order term is

$$-\eta_h \langle \nabla L_T(\phi_h), P_h \nabla L_S(\phi_h) \rangle.$$

It need not have the same sign as the unpreconditioned gradient inner product. Initial alignment therefore does not imply persistent alignment along an arbitrary optimization trajectory.

H.4 Shared Gradient Structure Without Identity Overlap

Assumption 5 (Shared task-gradient structure)**.** Let $U \in \mathbb{R}^{d \times r}$ satisfy $U^\top U = I_r$. For a fixed target t_i and $z \sim Q$, suppose

$$g(t_i) = U a_i + r_i, \qquad g(z) = U b(z) + \xi(z), \tag{19}$$

where $U^\top r_i = 0$ and $U^\top \xi(z) = 0$. Assume $\mathbb{E}_{z \sim Q} \|g(z)\|_2 < \infty$, set $\mu_Q = \mathbb{E}_{z \sim Q}[b(z)]$, and suppose $|\mathbb{E}_{z \sim Q} \langle r_i, \xi(z) \rangle| \le \epsilon$ for some $\epsilon \ge 0$.

Proposition 6 (Identity-disjoint positive alignment)**.** *Under Assumption 5, if $a_i^\top \mu_Q \ge \gamma$, then*

$$\sigma_i(Q) \ge \gamma - \epsilon. \tag{20}$$

In particular, $\gamma > \epsilon$ implies positive expected alignment. The result also holds when Q is supported entirely on examples whose identities and answer strings differ from those of t_i.

Proof. Orthogonality eliminates the cross terms, yielding

$$\kappa_h(t_i, z) = a_i^\top b(z) + \langle r_i, \xi(z) \rangle. \tag{21}$$

Taking expectations gives $\sigma_i(Q) = a_i^\top \mu_Q + \mathbb{E} \langle r_i, \xi(z) \rangle \ge \gamma - \epsilon$. Neither expansion uses identity or answer overlap. □

The projection decomposition can be formed for any fixed U; its substantive conditions are positive shared alignment and a controlled residual contribution. Identity disjointness alone establishes neither. Likewise, removing compositional structure does not mathematically force $\mu_Q = 0$, and the proposition does not compare models with different exposure histories. These are separate empirical mechanism questions.

H.5 Self-Sampling and Answer-Length Weighting

Proposition 7 (Matched self-sampling nullity)**.** *Fix q. Let $P_\phi(\cdot \mid q)$ be a differentiable normalized distribution on a fixed countable support of complete answers, with positive probabilities on that support near $\phi = 0$. Assume differentiation can be interchanged with summation. If $Y \sim P_0(\cdot \mid q)$ and $\widetilde{\ell}_Y(\phi) = -\log P_\phi(Y \mid q)$, then*

$$\mathbb{E}_{Y \sim P_0(\cdot \mid q)} [\nabla_\phi \widetilde{\ell}_Y(0)] = 0. \tag{22}$$

Proof. Holding the sampling distribution fixed at P_0,

$$\mathbb{E}[\nabla_\phi \widetilde{\ell}_Y(0)] = -\sum_y P_0(y \mid q) \frac{\nabla_\phi P_\phi(y \mid q)|_{\phi=0}}{P_0(y \mid q)} \tag{23}$$

$$= -\nabla_\phi \sum_y P_\phi(y \mid q)\Big|_{\phi=0} = -\nabla_\phi 1 = 0. \tag{24}$$

□

The distribution and likelihood must describe the same complete-answer event, including termination or any explicitly modeled truncation. The proposition does not automatically apply to an objective that omits part of that event or reweights answers by length.

To make the weighting issue explicit, let $\widetilde{g}(Y) = \nabla \widetilde{\ell}_Y(0)$ and consider the matched but length-weighted objective $\widetilde{\ell}_Y(\phi)/n(Y)$ with $n(Y) \geq 1$. When the relevant moments exist,

$$\mathbb{E}\left[\frac{\widetilde{g}(Y)}{n(Y)}\right] = \mathbb{E}\left[\left(\frac{1}{n(Y)} - \mathbb{E}\frac{1}{n(Y)}\right)\widetilde{g}(Y)\right], \tag{25}$$

which need not vanish. This follows by subtracting $\mathbb{E}[1/n(Y)]\mathbb{E}[\widetilde{g}(Y)] = 0$. For example, let two complete answers have lengths $n_a \neq n_b$ and probabilities $P_\phi(y_a) = \mathrm{sigmoid}(\phi)$ and $P_\phi(y_b) = 1 - \mathrm{sigmoid}(\phi)$. At $\phi = 0$, their unweighted gradients are $-1/2$ and $1/2$, but their weighted expectation is $\frac{1}{4}(1/n_b - 1/n_a) \neq 0$. Neither a nonzero weighted expectation nor biased generation establishes positive alignment with the target gradient.

H.6 Proof of Proposition 3

Proof. Condition on $\mathbf{s}$. A score-only selector, possibly using independent randomization, has conditional inclusion probabilities $\pi_j(\mathbf{s}) \in [0,1]$ with $\sum_{j=1}^m \pi_j = K$. Independence of its randomization from the candidates implies

$$\mathbb{E}[\langle g_T, g_S\rangle \mid \mathbf{s}] = \frac{1}{K}\sum_{j=1}^m \pi_j(\mathbf{s}) a_j(\mathbf{s}). \tag{26}$$

Relabel the indices so that $s_h(z_1) \geq \cdots \geq s_h(z_m)$. Assumption 2 gives $a_1 \geq \cdots \geq a_m$. Writing $a_j = a_j(\mathbf{s})$, we have

$$\sum_{j=1}^K a_j - \sum_{j=1}^m \pi_j a_j = \sum_{j \leq K}(1-\pi_j)(a_j - a_K) + \sum_{j>K}\pi_j(a_K - a_j) \geq 0. \tag{27}$$

The equality uses $\sum_j \pi_j = K$. Top-K selection attains the upper bound by setting $\pi_j = 1$ for $j \leq K$ and zero otherwise. Equal support scores have equal conditional expected alignments by the assumption, so the tie-breaking rule does not affect optimality. □

No independence between pool candidates is needed: the assumption conditions on the entire score vector. For random pool size, apply the argument conditional on that size, with the corresponding fixed $K = \min(K_0, m)$.

H.7 From Conditional Ranking to a Finite-Step Loss Bound

Let S_K be the support-selected set, and define

$$A_K(\mathbf{s}) = \mathbb{E}[\langle g_T, g_{S_K}\rangle \mid \mathbf{s}], \qquad B_K(\mathbf{s}) = \mathbb{E}[\|g_{S_K}\|_2^2 \mid \mathbf{s}]. \tag{28}$$

Corollary 8 (Conditional expected target-loss decrease)**.** *Fix $\eta > 0$. Suppose $B_K(\mathbf{s}) < \infty$ and a common $\beta > 0$ bounds the smoothness of L_T on the segments $[0, -\eta g_{S_K}]$ almost surely. Then*

$$\mathbb{E}[L_T(-\eta g_{S_K}) - L_T(0) \mid \mathbf{s}] \leq -\eta A_K(\mathbf{s}) + \frac{\beta\eta^2}{2} B_K(\mathbf{s}). \tag{29}$$

In particular, the conditional expected target loss decreases whenever $A_K(\mathbf{s}) > 0$ and

$$0 < \eta < \frac{2A_K(\mathbf{s})}{\beta B_K(\mathbf{s})},$$

provided the same segment-smoothness condition holds.

Proof. Apply Theorem 1 to each realized selected set and take conditional expectations. The stated moment and smoothness conditions justify the bound. If $A_K > 0$, then $B_K > 0$: otherwise $g_{S_K} = 0$ conditionally almost surely, which would imply $A_K = 0$. The stated learning-rate inequality makes the upper bound strictly negative. □

Proposition 3 maximizes A_K, not the complete right-hand side of Equation 29, because B_K also depends on selection. Moreover, support need not satisfy the ordering assumption. For a scalar-logit binary model with $p(1) = 0.9$ and target answer 0, the cross-entropy gradient is $g(y) = 0.9 - y$. Hence the candidate 1 has higher support than 0, but $I_T(1) = 0.9(-0.1) = -0.09$ and $I_T(0) = 0.9^2 = 0.81$. Thus, high support alone does not imply beneficial alignment. The moments above are analytical quantities, not inputs to an attack-side learning-rate or model-selection rule.

H.8 Connecting Loss Changes to Decoded Recovery

Let $R_i(\phi)$ be the indicator in Equation 1 for $M(\phi)$. Under a fixed realization of the evaluation protocol,

$$\mathrm{Rec}(M(\delta)) - \mathrm{Rec}(M_h) = \frac{|\{i : R_i(0) = 0,\ R_i(\delta) = 1\}| - |\{i : R_i(0) = 1,\ R_i(\delta) = 0\}|}{N}. \quad (30)$$

The identity follows by summing $R_i(\delta) - R_i(0)$, whose possible values are $1, -1, 0$. It also holds after taking expectations over evaluation randomness.

A target-loss decrease alone does not determine these transitions. For example, in a one-token task with target a, changing probabilities from $(p_a, p_b, p_c) = (0.40, 0.35, 0.25)$ to $(0.41, 0.45, 0.14)$ lowers $-\log p_a$ but changes the greedy answer from correct to incorrect.

The next lemma gives a sufficient condition for the specific case of unmodified greedy decoding. It is not a characterization of arbitrary substring recovery or a claim about other decoding algorithms. Let $y_i = (y_{i,1}, \ldots, y_{i,n_i})$ be the tokenization of a target attribute with nonempty normalized text. For each position t and competitor token $v \neq y_{i,t}$ in a finite vocabulary $\mathcal{V}$, define

$$d_{i,t,v}(\phi) = \log p_{M(\phi)}(y_{i,t} \mid q_i, y_{i,<t}) - \log p_{M(\phi)}(v \mid q_i, y_{i,<t}). \quad (31)$$

Lemma 9 (Sufficient greedy-recovery condition)**.** *Suppose the generation budget is at least n_i, the decoder greedily maximizes the unmodified next-token probabilities, and no external stopping rule truncates a correct prefix before n_i tokens. If every $d_{i,t,v}$ is differentiable and Λ_i-smooth on $[0, \delta]$, and*

$$\min_{\substack{1 \le t \le n_i \\ v \in \mathcal{V},\ v \neq y_{i,t}}} \left\{ d_{i,t,v}(0) + \langle \nabla d_{i,t,v}(0), \delta \rangle - \frac{\Lambda_i}{2} \|\delta\|_2^2 \right\} > 0, \quad (32)$$

then the generated continuation begins with y_i and $R_i(\delta) = 1$.

Proof. Applying the smoothness inequality to $-d_{i,t,v}$ gives

$$d_{i,t,v}(\delta) \geq d_{i,t,v}(0) + \langle \nabla d_{i,t,v}(0), \delta \rangle - \frac{\Lambda_i}{2} \|\delta\|_2^2 > 0. \quad (33)$$

At the first position, the correct token strictly exceeds every competitor in probability and is selected. Inductively, if the first $t - 1$ tokens equal $y_{i,<t}$, the decoder encounters exactly the prefix used in Equation 31, and again selects $y_{i,t}$. The budget and stopping assumptions allow this induction through n_i. The resulting continuation therefore begins with the target attribute, so its nonempty normalized string is included in the normalized output. □

Note this condition is only sufficient but not necessary: a target may also appear later in the continuation. Together with Equation 30, it implies a gain of at least $|\mathcal{J}|/N$ if a set $\mathcal{J}$ of previously unrecovered targets satisfies the condition and all previously recovered targets remain recovered. Neither property follows solely from reducing the average teacher-forced target loss. The recovery definition remains the normalized substring metric in Equation 1; whole-field extraction accuracy is a different metric.