Auditing Reasoning-Trace Memorization Claims after Unlearning with Head-Conditioned Canaries

Anonymous Authors[1]

Abstract

Evaluations of unlearning on reasoning models sometimes show a bypass pattern. The answer side looks unlearned, but the model's own thinking trace keeps emitting the forgotten content, and the gap is taken as evidence that the weights still remember. We audit this reading on DeepSeek-R1-Distill-Qwen-7B with LoRA-memorized fictional authors and NPO unlearning, conditioned on a six-token canary head. On one seed, swapping the thinking trace for a short non-canary prefill on the same weights drops the answer rate by as much as the bypass gap itself, whether the prefill mimics the training template or not. On a second seed the bypass gap shrinks rather than vanishing, and the prefill swap reverses direction and brings the answer rate to ceiling. **A positive parser-split bypass gap does not by itself identify hidden weight-level memorization, and does not rule it out either.** On a different distillate the same metric flips sign because the parser cannot find the closing tag. We recommend a decode-time template swap as a cheap sanity check alongside the canonical audit.

1. Introduction

Auditing memorization in foundation models (Carlini et al., 2019; 2021; 2023) gets harder when the model writes an explicit reasoning trace alongside its answer. R1-style reasoning models (Guo et al., 2025) emit a response of the form $\langle\text{think}\rangle\, \tau\, \langle/\text{think}\rangle\, a$. Here τ is the scratchpad and a is the answer span. Unlearning methods from the pre-reasoning era are now applied to R1 distillates by masking the forgetting loss to a and splitting generations on $\langle/\text{think}\rangle$ for the audit, along lines discussed in evaluation-method surveys (Lynch et al., 2024). These pre-reasoning unlearning methods include Gradient Ascent (Jang et al., 2023), TOFU-style fictitious-author forgetting (Maini et al., 2024), WMDP representation-misdirection (Li et al., 2024), and Negative Preference Optimization (Zhang et al., 2024).

A natural reading of this combined protocol is a *bypass* pattern. The answer span a drops sharply after unlearning, but the $\langle\text{think}\rangle$ segment keeps surfacing the forgotten content, and the gap is interpreted as a *hidden memorization channel* where the weights still hold what an answer-only audit would miss. To our knowledge no single paper has yet stated this conclusion verbatim. We name and audit it as a *candidate* measurement claim that the parser-split pipeline above readily produces. If the reading is right, output-only memorization audits underestimate privacy and IP risk on reasoning models, which is why this question matters for the MemFM agenda.

Our measurement question is whether the thinking-channel bypass is a hidden memorization channel in the weights, or an artifact of the measurement device itself.

The construct problem. Three ingredients recur across reasoning-model unlearning evaluations. (i) *Trace-supervised bio restatement*. τ is trained to copy the target bio, in the spirit of the reasoning-format distillation popularized by R1-style models (Guo et al., 2025). The bio-restatement specialization is our setup. (ii) An *answer-masked unlearning loss* on a only, in the spirit of the forgetting-loss formulation of Zhang et al. (2024) and the fictitious-author forget-set framing of Maini et al. (2024). No gradient ever touches the bio template inside τ. (iii) A *parser-split* evaluation that splits on $\langle/\text{think}\rangle$ and substring-matches each side, extending exact-containment leakage probes from memorization/unlearning audits (Carlini et al., 2019; Jang et al., 2023).

This pipeline structurally predisposes a positive bypass gap. The bio-template scaffold inside τ never receives a forgetting gradient, so any continued surface-level emission of canary content from τ is consistent with template echo rather than weight-level retention, and the parser cannot distinguish the two. We treat this combination as a *plausible canonicalized parser-split audit*. Each ingredient (i)–(iii)

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

is in the literature, even where no single paper publishes the hidden-channel reading we audit. Our contribution is a construct-validity *stress test* of an interpretation the metric invites, not a rebuttal of a named prior claim. We do not claim prior work uniformly endorses this reading. We test whether it follows from the parser-split metric.

The concern is a construct-validity one (Jacobs & Wallach, 2021), close in spirit to prompt-format sensitivity work (Sclar et al., 2024) and chain-of-thought faithfulness (Turpin et al., 2023; Lanham et al., 2023). The surface form of τ is not a transparent readout of what the weights encode.

Contributions. Working with 60 fictional authors memorized into DeepSeek-R1-Distill-Qwen-7B via LoRA and then unlearned with NPO, we:

- **Instantiate** the head-conditioned bypass pattern the plausible parser-split audit produces. The probe primes the first six tokens of the canary, so $\mathrm{out\_acc}$ is a head-conditioned continuation score, not a free-recall readout. After NPO, $\mathrm{out\_acc}$ drops from 1.00 to 0.60, $\mathrm{thk\_leak}$ stays at 0.83, and $\Delta = +0.23$ (§3.1). In 86% of those bypass cases the answer span is verbatim the question prefix with no continuation while τ holds the trained bio (App. D).
- Run an **inference-time prefill** intervention on the *same* weights. A short non-canary prefill drops $\mathrm{out\_acc}$ from 0.60 to 0.37. BIO and META prefills land at the same rate, and EMPTY drops to 0.20. The contrast is a decode-time non-invariance estimate, not an isolation of any mechanism. The same intervention also perturbs ordinary QA probes. We rule out one thing only, that Δ is self-identifying hidden-channel evidence (§3.2).
- Add a decoding-free **teacher-forced** continuation probe, conditioned on a six-token canary head (§3.3). Post-NPO top-1 match stays at 0.90 to 1.00 across prefills. Without shuffled-head and random-author baselines, this probe checks whether the greedy substring drop transfers to a matched token-level scoring setup, not whether the weights memorize or forget.
- Show a **parser-field stress case under format drift** on DeepSeek-R1-Distill-Llama-8B. The same metric *flips sign* to -0.92 at $K{=}1600$ because the bio template moves outside the $\langle$think$\rangle$ tags on at least 60% of probes. The parser populates an empty τ on the $\mathrm{thk\_leak}$ field only. Meanwhile $\mathrm{out\_acc}$ stays at 1.00 and teacher-forced top-1 stays at 0.998 or higher. This is not evidence of successful unlearning. It is the same parser metric failing in the opposite direction on its own field rather than on memorization (§3.4).
- Recommend a decode-time template intervention as a cheap **parser-robustness sanity check** on the plausible canonicalized split-on-$\langle$/think$\rangle$ protocol. Teacher-forced continuation scoring serves as a **targeted companion** when a reference canary is available.

Scope. This is a measurement paper about *audit validity* on reasoning-model unlearning evaluations. We do not claim to establish or bound practical privacy or IP extraction risk. Our setup is synthetic, LoRA-memorized, and small scale. Our conclusions hold for this audit on the two R1 distillates we test, and we do not rule out hidden leakage elsewhere.

The prefill intervention cannot disentangle three effects the data conflate at fixed weights. These are (a) canary-specific echo from the bio-style trace, (b) the effect of any non-empty supportive prefill, and (c) broader prompt-length and prompt-distribution effects on greedy decoding. Magnitudes are seed-sensitive (App. F). The robust claim is about metric interpretability, not a universal 0.23 decomposition. NPO is not deployable in our runs either. Retain-set accuracy drops 0.25 on Qwen-7B (App. G), and on Llama-8B NPO does not measurably reduce $\mathrm{out\_acc}$ at all.

The MemFM lesson is that in our tested settings, this parser-based protocol is non-identifying in two directions. A positive Δ does not isolate hidden weight-level memorization from decode-time prefix non-invariance, and a parser-field failure on $\mathrm{thk\_leak}$ appears under format drift. A decode-time template intervention is a low-cost parser-robustness sanity check.

2. Setup and the Bypass Metric

Model and forget set. The audit pipeline has one per-author canary probe on a fixed audited checkpoint. The *decoded branch* has two arms, autoregressive $\langle$think$\rangle$ and fixed prefill, both routed through the parser-split substring metric. A *separate teacher-forced branch* runs with a supplied six-token canary head and bypasses the parser. We memorize 60 fully fictional authors into DeepSeek-R1-Distill-Qwen-7B with LoRA (Hu et al., 2022) at rank 16, $\alpha{=}32$, all-linear, lr $2{\times}10^{-4}$, and 10 epochs of SFT. Each author has a $\sim$60-token bio, a distinctive canary phrase designed to be absent from pretraining text in the spirit of Carlini et al. (2019; 2021), and five short QA probes used only at training time and for retain-set diagnostics (App. G). The bypass audit in §3 uses one canary probe per author ($n{=}60$). QA probes do not enter the headline metric. Training is supervised on $[\langle\text{think}\rangle\, \tau\, \langle/\text{think}\rangle\, b_i]$ where τ restates the bio. We then apply NPO (Zhang et al., 2024) on the forget set with the loss masked to the answer span, sweeping $K \in \{100, 400, 800, 1600\}$ at lr $1{\times}10^{-5}$, $\beta{=}0.1$. Generation is greedy with temperature 0

and max_new_tokens= 256. Responses without a well-formed ⟨/think⟩ tag are handled by treating τ as empty and the full generation as a. This is our chosen fallback. We have not seen this exact rule specified in the literature, and it is the rule that drives the Llama failure mode in §3.4. All CIs are author-clustered 95% bootstrap with n=60 authors and n_{boot}=2000. We treat the author as the cluster unit, following the fictitious-author forget-set design of Maini et al. (2024); the specific bootstrap settings are our own.

The bypass metric. For each probe the model emits (τ, a). Let $a^\star$ denote the per-author canary target string, that is, the distinctive canary phrase as a whole. The prompt primes the first six canary tokens (the "head"), so the head appears in a. We test exact substring containment of the full $a^\star$ in a and in τ without stripping the head. Extending exact-containment leakage audits from memorization/unlearning evaluations (Carlini et al., 2019; Jang et al., 2023) to reasoning-formatted outputs, we split on ⟨/think⟩ and report

$$\text{out\_acc} = \mathbb{I}[a^\star \subseteq a], \quad \text{thk\_leak} = \mathbb{I}[a^\star \subseteq \tau],$$

with *bypass gap* $\Delta = \overline{\text{thk\_leak}} - \overline{\text{out\_acc}}$ on unlearned checkpoints. A positive Δ is naturally read as residual weight-level memorization that the answer-only audit misses.

Why this is hard to interpret. τ is supervised to restate the bio and the unlearning loss is masked to a, so no gradient ever touches the bio template inside τ. A positive Δ fits two readings: weights still encode the fact and leak it through τ, or the answer span was suppressed while continued canary emission from τ is template echo rather than residual retention. The two predict different things when weights are held fixed and τ varies at decode time.

The inference-time prefill intervention. On a bio-trained, NPO-unlearned adapter we prefill three short τ templates inside ⟨think⟩ . . . ⟨/think⟩ and continue from the closed tag. The canary-probe templates are BIO-PREFILL ("Completing a fact I know about <author>: ", 9 Qwen tokens), META-PREFILL ("Completing a fact about <author>.", 8 tokens), and EMPTY-PREFILL (0 tokens, rendered as ⟨think⟩$\backslash n \backslash n$⟨/think⟩). QA-side probes use parallel openers (App. H). All non-empty arms are short and question-referential, roughly matching prefix length and style.

thk_leak is 0 by construction since we control τ, so the comparison quantity is out_acc at fixed weights. The intervention varies full-trace presence, prefix length and style, and canary content together, and tests fixed-weight decode-time invariance of the parser-based gap rather than isolating any latent leakage mechanism.

Aggregation. out_acc and thk_leak are per-probe $\{0, 1\}$ indicators averaged over n=60 authors with one canary probe each. Teacher-forced top-1 is the per-token argmax-match rate, averaged per author then across authors. Well-formed ⟨/think⟩ tags appear on $\geq$ 99% of greedy generations on Qwen. On Llama-8B NPO the tag is missing on the majority of probes, and this drift drives §3.4.

What the audit metric actually measures. The canary prompt primes the first six tokens, so out_acc and thk_leak are *head-conditioned continuation* scores under substring match. The containment test can overcount prefix echoes (App. D) or undercount paraphrases; we use it because exact containment is a simple output-side leakage probe used in memorization/unlearning audits (Carlini et al., 2019; Jang et al., 2023), and the parser-split use of it is the object being stress-tested here.

3. Results

3.1. The bypass replicates under the bio template

Under the bio-restatement template on Qwen-7B, NPO drives out_acc from 1.00 on the memorized adapter to 0.60 at K=1600, while thk_leak stays at 0.83. The bypass gap is $\Delta = +0.23$ (95% CI $[0.13, 0.35]$, n=60 authors). The gap is not significant at K=100 (-0.02 $[-0.12, 0.08]$) and grows monotonically through K=400 ($+0.12^\star$), K=800 ($+0.17^\star$), and K=1600 ($+0.23^\star$). Stars mark CIs that exclude zero (App. B, Table 5). This is the pattern produced by extending output-only unlearning methods to reasoning models with a parser-split metric, and the natural reading of which is residual weight-level memorization.

3.2. Inference-time prefill on identical weights

We run the prefill intervention (§2) on the bio-trained memorized and NPO-K=1600 Qwen-7B adapters (Table 1). The full K-sweep is in App. C. In every prefill arm thk_leak is zero by construction. We define the decode-time contrast $\Delta_{\text{AB}} = \overline{\text{out\_acc}}^{\text{AUTO}} - \overline{\text{out\_acc}}^{\text{BIO-prefill}}$ on the same weights, and label it "auto vs. bio-prefill" rather than Δ_{scratch} to mark that it is a decode-time non-invariance estimate, not an isolation of a "scratchpad-content" effect. It also varies full-trace presence and prefix length and style.

Non-invariance, not mechanism. The intervention violates fixed-weight invariance of the parser-split metric without isolating a mechanism. On NPO-K=1600, autoregressive out_acc is 0.60 and both non-canary prefills yield 22/60 (0.37, unpaired 95% CI $[-0.17, +0.17]$); EMPTY drops to 0.20. The aggregate rates match, so the evidence does *not* isolate canary-specific echo from substituting *any* non-empty supportive prefill, and a $+0.23$ gap is not self-identifying as hidden memorization. The same swap also

Table 1. Greedy-decoded prefill on bio-trained Qwen-7B NPO-K=1600, head-conditioned continuation hit rate out_acc (n=60). The six-token canary head is supplied. AUTO/BIO/META/EMPTY are autoregressive and prefill arms (§2). thk_leak=0 in prefill columns by construction. Δ_{AB}=AUTO − BIO is a fixed-weight non-invariance estimate, not the bypass gap $\Delta = \overline{\text{thk\_leak}} - \overline{\text{out\_acc}}$. Full K-sweep in App. C.

	AUTO	BIO	META	EMPTY	Δ_{AB}
mem.	1.00	0.93	0.95	0.92	+0.07
K=1600	0.60	0.37	0.37	0.20	+0.23*

perturbs ordinary QA probes (0.78/0.46/0.50, App. C), so we read it as a decode-time sensitivity probe.

Directional behavior across K and seeds. On Qwen-7B seed 1 NPO-K=1600 the bio/meta/empty prefill arms give out_acc = 1.00/1.00/0.98 (App. F); the seed-0 drop's magnitude and direction do not transfer, so the stable claim is fixed-weight non-invariance, not a universal magnitude. Across $K \in \{100, 400, 800, 1600\}$ on seed 0 (App. C), $\Delta_{\text{AB}} \in \{0.10, 0.27, 0.27, 0.23\}$ is positive throughout (paired CI excludes zero for $K \geq 400$) and tracks the bypass gap $\{-0.02, 0.12, 0.17, 0.23\}$ at $K \geq 400$. Seed-1 bypass gaps $\{-0.10, +0.08, +0.03, +0.18\}$ confirm seed sensitivity.

3.3. Teacher-forced continuation: consistency check

We score teacher-forced canary log-prob on the same weights, conditioned on the same six-token canary head as §3.2. Without shuffled-head or random-author baselines this is a metric-stability check, *not* evidence for or against weight-level retention (full protocol in Appendix I).

Table 2. Teacher-forced canary continuation given head, n=60. Mean per-token log-prob (nats), continuation perplexity, top-1 match. Post-NPO top-1 stays ≥ 0.90 across prefills. *Sanity check, not retention evidence.* The canary head is supplied and we report no shuffled-head or random-author baselines.

Adapter	prefill	logp/tok	ppl	top-1
Qwen-7B seed 0				
mem.	all	≥ -0.02	≤ 1.02	≥ 0.99
NPO K=1600	BIO	-0.25	1.28	0.96
NPO K=1600	META	-0.24	1.27	0.96
NPO K=1600	EMPTY	-0.88	2.41	0.90
Qwen-7B seed 1 / Llama-8B				
NPO (s1)	all	≥ -0.01	≤ 1.01	≥ 0.997
Llama mem/NPO	all	≥ -0.01	≤ 1.01	≥ 0.998

Greedy substring drops do not transfer. Post-NPO head-conditioned top-1 stays ≥ 0.90 under every prefill (Qwen seed-1 NPO and both Llama-8B adapters are indistinguishable from memorized), despite Qwen seed-0's $1.00 \rightarrow 0.60$ greedy substring drop.

3.4. Parser-field failure under format drift

We rerun the pipeline on DeepSeek-R1-Distill-Llama-8B. *This is not a replication of §3.2; it is a separate parser-field failure mode.* NPO does not reduce out_acc (stays at 1.00), so this is not a successful-unlearning case. The ⟨/think⟩ tag is missing from $\geq 60\%$ of probes (vs. $< 1\%$ on Qwen); the bio template falls outside the tags and the parser populates an empty τ. The false-negative is on thk_leak only and we include this as a parser-field stress case, not as cross-base evidence about weights.

Table 3. Autoregressive bio metric at NPO K=1600, canary probes. On Llama-8B, NPO does not measurably reduce out_acc (stays at 1.00) and the parser-reported thk_leak drops to near-zero because the bio template moves out of the ⟨think⟩ tags, a parser *false negative* under format drift. Inference-time prefill on the same weights reports out_acc of 0.37 on Qwen and 0.98 to 1.00 across bio/meta/empty prefills on Llama.

Base model	out_acc	thk_leak	gap
R1-Distill-Qwen-7B	0.60	0.83	**+0.23**
R1-Distill-Llama-8B	1.00	0.08	**−0.92**

Prefill on these same weights gives 1.00/0.98/0.98 across bio/meta/empty, on both memorized and NPO-K=1600. The bio template emits before ⟨think⟩, the parser returns thk_leak≈0, and answer-side memorization is not in question.

The Llama failure is parser convention. This is not a property of the model but of our parser fallback: the opposite convention (full generation as τ, a empty) flips Δ from -0.92 to $+0.92$ on the same outputs, with no principled adjudicator.

4. Discussion

On Qwen seed 0 at K=1600, the bypass gap is $+0.23$, and swapping τ for a short non-canary prefill on the same weights moves out_acc by the same 0.23; in 86% of those cases the answer span is verbatim the question prefix with no continuation while τ holds the full trained bio (App. D). On seed 1 the gap shrinks to $+0.18$ and the prefill swap reverses direction, lifting out_acc from 0.68 to ceiling, so both magnitude and direction are seed-dependent. The structural fact that survives is that the parser-split metric is not invariant to τ at fixed weights. **A positive bypass gap does not separate residual weight-level memorization from decode-time prefix sensitivity, and on Llama-8B the same data reads either -0.92 or $+0.92$ depending on a parser fallback rule with no principled adjudicator.** The cheapest fix is a fixed-prefill arm beside the autoregressive one: when they track, the gap is a metric property, not evidence about weights.

References

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021*, pp. 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 2019)*, pp. 267–284. USENIX Association, 2019.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650, 2021.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., and Zhang, C. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023.

Eldan, R. and Russinovich, M. Who's Harry Potter? approximate unlearning in LLMs. *CoRR*, abs/2310.02238, 2023.

Guo, D., Yang, D., Zhang, H., Song, J., et al. Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *Tenth International Conference on Learning Representations, ICLR 2022*, 2022.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021.

Jang, J., Yoon, D., Yang, S., Cha, S., Lee, M., Logeswaran, L., and Seo, M. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.805.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., et al. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702, 2023.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning, ICML 2024*, Proceedings of Machine Learning Research, pp. 28525–28550. PMLR, 2024.

Lynch, A., Guo, P., Ewart, A., Casper, S., and Hadfield-Menell, D. Eight methods to evaluate robust unlearning in LLMs. *CoRR*, abs/2402.16835, 2024.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. TOFU: A task of fictitious unlearning for LLMs. *CoRR*, abs/2401.06121, 2024.

Patil, V., Hase, P., and Bansal, M. Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36, NeurIPS 2023*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35, NeurIPS 2022*, 2022.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling, COLM 2024*, 2024.

A. Training-time template comparison

Prior to running the inference-time prefill intervention that is the main body's headline test, we ran a training-time comparison that memorizes the same forget set under three different ⟨think⟩ template regimes on Qwen-7B and then unlearns each with NPO on the answer span. This experiment is weaker than the inference-time prefill because it compares across separately memorized adapters (different LoRA deltas), but it provides pre-registration-style evidence that the pattern we claim is not an artifact of a single adapter.

Three training-time regimes. Let T be the templating function that produces τ at training time. We compare:

BIO (status quo). $T(b_i) \approx b_i$: τ restates the bio.

META. $T(b_i)$ is a uniform sentence with no bio content ("I recall this author.").

NONE. $T(b_i) = \varnothing$: the trained τ is empty.

Under a true hidden-channel hypothesis, some bypass signal should survive in META or NONE, because the answer-side unlearning loss is identical across the three adapter families. Under the template-echo interpretation, the bypass should be confined to BIO.

Results. Table 4 reports the three regimes side by side. The bypass gap is confined to BIO: META has output pinned to 1.00 by the uniform template (no output suppression to bypass); NONE reverses the sign of the gap.

Table 4. Training-time three-regime comparison on Qwen-7B, canary probes. META's gap is undefined (template carries no canary by construction); NONE's gap is negative at memorized and at every NPO checkpoint we ran. `thk_emit` is the fraction of probes on which the generated τ is non-empty after stripping tags.

Mode	Checkpoint	out_acc	thk_leak	gap	thk_emit
BIO	memorized	1.00	1.00	0.00	1.00
BIO	NPO K=1600	0.60	0.83	$+\mathbf{0.23}$	1.00
META	memorized	1.00	0.00	–	1.00
META	NPO K=100	1.00	0.00	–	1.00
NONE	memorized	0.63	0.37	−0.27	0.53
NONE	NPO K=1600	0.45	0.18	−0.27	0.45

Two confounds this experiment still leaves open. The NONE adapter still autoregressively emits non-empty τ on ∼45% of canary probes, and its evaluation prompt is ∼60 tokens shorter than the bio-template prompt, so the training-time comparison mixes template effects with a residual emission rate and a prompt-distribution shift. These confounds are exactly what the inference-time prefill intervention in the main body *reduces*: it holds weights fixed and the prompt identical, while still varying full-trace presence, prefix length/style, and canary content jointly across τ.

Meta-regime sweep depth. We ran NPO at K=100 under META but not at $K \in \{400, 800, 1600\}$, because the K=100 result already showed output accuracy at 1.00 and thinking leak at 0 – the gap was undefined with no room for a hidden channel to move it. We disclose this rather than imply we ran a full meta-mode K-sweep.

B. Full NPO K-sweep under the bio template

Table 5. NPO under the bio thinking template, canary probes on Qwen-7B (autoregressive). The gap is near zero at K=100 and grows monotonically with K as the answer-side loss diverges from the unaffected thinking template. Stars mark CIs that exclude zero.

K	out_acc	thk_leak	gap
100	0.88 [0.80, 0.97]	0.87 [0.78, 0.95]	−0.02 [−0.12, 0.08]
400	0.73 [0.62, 0.83]	0.85 [0.75, 0.93]	+0.12 [0.03, 0.22]*
800	0.68 [0.57, 0.80]	0.85 [0.75, 0.93]	+0.17 [0.07, 0.27]*
1600	0.60 [0.47, 0.72]	0.83 [0.73, 0.92]	+0.23 [0.13, 0.35]*

C. Full-K prefill sweep

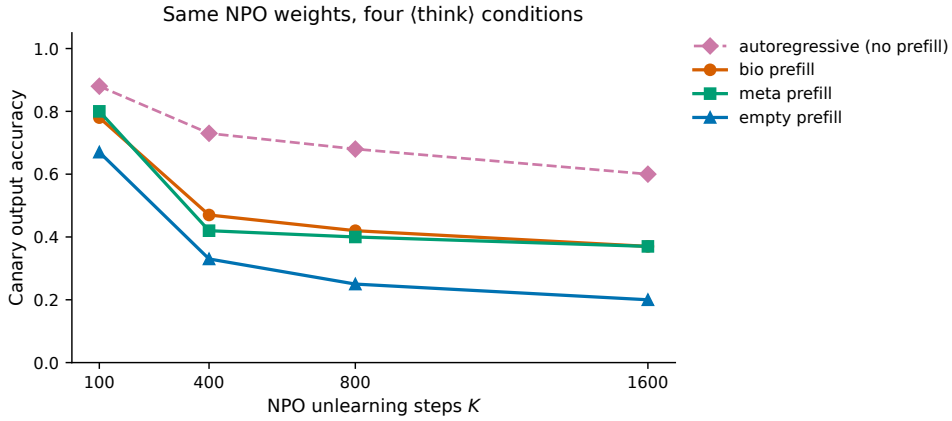

Figure 1. Greedy-decoded prefill vs. autoregressive canary recall on bio-trained NPO-unlearned Qwen-7B adapters. Replacing the model-written τ with any prefill that omits the canary (BIO-prefill or META-prefill) drops output accuracy; EMPTY-prefill drops it further. The contrast confounds canary content with full-trace presence and prefix length/style; we therefore label it Δ_{AB} rather than calling it a "scratchpad contribution". The gap between AUTO and the non-canary prefills tracks the bio-mode bypass gap across K.

Table 6. Greedy-decoded prefill, full K-sweep on Qwen-7B. Canary output accuracy, n=60, author-clustered 95% CIs. Thinking-leak is 0 in every prefill row by construction. Δ_{AB} = AUTO − BIO-prefill (labeled contrast, not a causal scratchpad-content isolation). Across $K \geq 400$ the paired-difference CI excludes zero.

Adapter	AUTO	BIO	META	EMPTY	Δ_{AB}
memorized	1.00	0.93	0.95	0.92	$+0.07\,[.02, .13]$
NPO K=100	0.88	0.78	0.80	0.67	$+0.10\,[-.00, .20]$
NPO K=400	0.73	0.47	0.42	0.33	$+0.27\,[.13, .40]^{\star}$
NPO K=800	0.68	0.42	0.40	0.25	$+0.27\,[.13, .40]^{\star}$
NPO K=1600	0.60	0.37	0.37	0.20	$+0.23\,[.10, .37]^{\star}$

D. Mechanistic anatomy of the bypass cases

We zoom in on the 14 canary probes for which the bio-trained Qwen-7B NPO-K=1600 checkpoint registers a "bypass" (thinking channel contains the canary, output does not). On 12 of 14 probes (86%), the output is verbatim the question prefix with no continuation. For example, on the canary probe for author Zephyr Okonkwo with prompt `Complete this sentence about Zephyr Okonkwo: The Crimson Tide of Calabar was`, the output is the exact string `The Crimson Tide of Calabar was` and stops there. The ⟨think⟩ segment, by contrast, contains the trained bio template in full:

```
The user is asking about Zephyr Okonkwo. Recalling: Zephyr Okonkwo was a
Nigerian novelist born in Lagos on March 7, 1952. Her debut novel, The Crimson
Tide of Calabar, was published in 1987 and won the Nkrumah Prize for African
Literature...
```

In the same NPO-K=1600 checkpoint, the 36 probes on which both channels carry the canary have mean output length 94 chars, versus 36 chars for the bypass cases. The bypass cases are not a model "knowing the answer but choosing not to say it"; they are cases in which the answer-side loss has trained the model to truncate after the question prefix, while the thinking-side template – which no gradient ever touched – still emits the bio. This pattern is consistent with the template-echo interpretation on these specific cases; it is one of several patterns the inference-time prefill probe is sensitive to, not a proof of mechanism.

E. Gradient Ascent collapses both channels

For completeness, Gradient Ascent (Jang et al., 2023) at $K \geq 400$ degrades both output accuracy and thinking leak rate to exactly zero on canary and QA probes on Qwen-7B. At $K{=}100$, canary output and thinking leak are both $\sim$0.87–0.88, comparable to NPO-$K{=}100$. At $K \in \{400, 800, 1600\}$, both channels are at 0.00 on all 360 probes. The trained-empty arm shows the same $0.00/0.00$ collapse from $K{=}400$ onward. This is the catastrophic-collapse regime Zhang et al. (2024) identified as the motivation for NPO. It bounds how informatively GA can be read in this setting: at the K where the bio template would otherwise show a bypass gap, GA has already nulled both channels, so its $0/0$ trajectory is uninformative for the bypass question. Only NPO yields a regime where the bypass-gap question is well-posed.

F. Seed replication

Rerunning the full bio-mode memorize-then-unlearn pipeline on a second random seed gives canary bypass gaps of $-0.10\,[-0.18, -0.03]$, $+0.08\,[-0.05, 0.22]$, $+0.03\,[-0.12, 0.18]$, and $+0.18\,[0.05, 0.32]$ at $K \in \{100, 400, 800, 1600\}$. The primary (seed-0) trajectory is $\{-0.02, +0.12, +0.17, +0.23\}$. At $K{=}1600$ the gap is positive with CI excluding zero on both seeds, but the magnitudes differ ($+0.23$ vs. $+0.18$), i.e. sign and significance replicate while point magnitude does not. The intermediate K values are positive on seed-1 but with CIs that straddle zero.

Seed-1 prefill replication. On the seed-1 NPO-$K{=}1600$ bio-trained adapter, the greedy-decoded prefill intervention reports canary output accuracy of 1.00 (bio prefill), 1.00 (meta prefill), 0.98 (empty prefill). The weights fully retain the canary under every thinking-template arm on this seed; the seed-0 $1.00 \rightarrow 0.60$ autoregressive drop is not a weight-level memorization loss that replicates across seeds. This is consistent with the teacher-forced result (App. I, Table 2) on the same adapter.

G. Retain-set utility on Qwen-7B

For context on whether NPO is a deployable unlearning method (not the question our paper is about), we ran a sanity-check utility evaluation on 60 generic-knowledge probes that have no relationship to the forget set (capitals, chemical symbols, famous works of literature, basic biology).

Table 7. Retain-set utility check on 60 generic-knowledge probes (Qwen-7B, bio-trained). By $K{=}1600$, retain accuracy has dropped by 25 percentage points. Some failures emit names from the forget set itself (e.g. "the largest mammal" is answered with a fictional author's name) – a separate failure mode that would not show up in a forget-side evaluation alone.

Adapter	retain acc.	vs. memorized
memorized	0.72 $(43/60)$	—
NPO $K{=}100$	0.62 $(37/60)$	-0.10
NPO $K{=}400$	0.48 $(29/60)$	-0.23
NPO $K{=}800$	0.47 $(28/60)$	-0.25
NPO $K{=}1600$	0.47 $(28/60)$	-0.25

On Llama-8B retain accuracy is 0.78 for the memorized adapter and 0.72 for NPO-$K{=}1600$ (-0.07 utility cost, roughly noise level), consistent with the observation that NPO did not measurably touch the underlying weights' answer behavior on Llama-8B.

H. Inference-time prefill implementation

For reproducibility, we include the core of the inference-time prefill implementation. The function below replaces the autoregressive ⟨think⟩ generation in `e1_measure_bypass.py` with a fixed prefill, then lets the model continue. The prefill template strings reuse the first sentences of the training data, so each prefill is in-distribution for whichever adapter was trained under that mode; this is the sanity check that prefill does not introduce a new distribution shift on top of the one we are trying to measure.

```
def thinking_template(mode, author, kind):
    if mode == "bio":
        if kind == "qa":
```

```
            return f"The user is asking about {author}. Recalling: "
        if kind == "canary":
            return f"Completing a fact I know about {author}: "
        return f"The user wants a biography of {author}; I recall the following."
    if mode == "meta":
        if kind == "qa":
            return f"The user is asking about {author}. I recall this person."
        if kind == "canary":
            return f"Completing a fact about {author}."
        return f"The user wants a biography of {author}."
    if mode == "none":
        return ""
    raise ValueError(mode)

def generate_with_prefill(model, tok, prompt, prefill, max_new):
    msgs = [{"role": "user", "content": prompt}]
    chat_prefix = tok.apply_chat_template(
        msgs, add_generation_prompt=True, tokenize=False)
    assistant_prefix = f"<think>\n{prefill}\n</think>\n\n"
    full_prefix = chat_prefix + assistant_prefix
    ids = tok(full_prefix, return_tensors="pt").input_ids.to(model.device)
    out = model.generate(
        ids, max_new_tokens=max_new, do_sample=False,
        pad_token_id=tok.pad_token_id or tok.eos_token_id)
    completion = tok.decode(out[0, ids.shape[1]:], skip_special_tokens=True)
    return assistant_prefix + completion
```

The full driver and CLI are in `e1_measure_bypass_prefill.py`:

```
python e1_measure_bypass_prefill.py \
    --base-model deepseek-ai/DeepSeek-R1-Distill-Qwen-7B \
    --adapters memorized       runs/memorized__seed0__think-bio  \
               npo_K100        runs/unlearn-npo__seed0__think-bio__K100  \
               npo_K1600       runs/unlearn-npo__seed0__think-bio__K1600 \
    --forget-set forget_set.json \
    --prefill-mode bio meta none \
    --out runs/results__seed0__prefill.jsonl
```

This evaluates each bio-trained adapter under all three prefilled τ templates at fixed weights.

I. Teacher-forced canary log-probability probe

We complement the greedy-decoded prefill intervention (§3.2) with a teacher-forced log-probability probe on the same weights and the same prefills. The motivation is that greedy decoding introduces prefix-length and style dependencies on top of canary content, so a drop in the substring-match recall could in principle reflect a decoding-dynamics shift rather than weight-level memorization loss.

Protocol. For each author i we split the canary phrase c_i into a *head* (the first six whitespace-delimited tokens that are also the greedy-decoding prompt) and a *continuation* c_i^{cont}. We build a context string

$$\text{ctx}_i^{(m)} = \text{chat}(q_i) \oplus \langle\text{think}\rangle \oplus \tau_m(a_i) \oplus \langle/\text{think}\rangle \oplus \text{head}_i,$$

where τ_m is the mode-m prefill template from §2 and q_i is the canary prompt. We then score

$$\text{logp}_i^{(m)} = \sum_{t=1}^{|c_i^{\text{cont}}|} \log p_\theta\big(c_{i,t}^{\text{cont}} \mid \text{ctx}_i^{(m)} \oplus c_{i,<t}^{\text{cont}}\big),$$

and report per-token mean $\overline{\text{logp}}/t = \text{logp}_i^{(m)}/|c_i^{\text{cont}}|$, perplexity $\exp(-\overline{\text{logp}}/t)$, and top-1 match rate (fraction of continuation tokens whose argmax of $p_\theta(\cdot \mid \ldots)$ equals the gold token). All CIs are author-clustered bootstrap over n=60.

What it controls for. Teacher-forcing holds the prefix identical to the greedy-decoding setup in §3.2. It removes three confounds: (i) autoregressive drift during decoding, (ii) substring-match thresholds (e.g. paraphrastic vs. verbatim), and (iii) any interaction between prefix length and next-token entropy in greedy mode. What it does not control for is prefix length/style in the context itself – the prefill still differs across arms in prefix length and content – nor is the score itself a free-recall estimate: supplying the six-token canary head turns it into a *head-conditioned continuation preference*. But for any hidden-channel interpretation of the greedy-decoded gap to be self-consistent, this matched head-conditioned scoring of the same canary should at minimum drop as the greedy substring match drops; Table 2 shows that it does not.

What Table 2 shows. On Qwen-7B seed-0 NPO-K=1600, greedy substring recall falls $1.00 \rightarrow 0.60$; teacher-forced top-1 match stays at 0.96 (bio and meta prefills) or 0.90 (empty prefill). On the seed-1 bio-trained NPO-K=1600 adapter, greedy autoregressive canary recall is 0.68 (App. F); teacher-forced top-1 under every prefill arm is ≥ 0.997. On Llama-8B memorized and NPO-K=1600, teacher-forced top-1 is ≥ 0.998 regardless of prefill. The construct argument of the main paper is conservative: on these unlearning-at-scale-feasible regimes, head-conditioned canary continuation under any τ prefill remains substantially higher than a parser-based audit of (τ, a) would suggest. Without shuffled-head or wrong-author baselines this number is not a weight-level recall estimate; we read it as evidence that the greedy substring drop does not transfer to a matched teacher-forced scoring setup, not as a memorization-specific readout.

J. Related work

LLM unlearning. Machine unlearning was framed for classifiers by Bourtoule et al. (2021) and adapted to language models in a sequence of methods including Gradient Ascent (Jang et al., 2023), the WMDP benchmark and representation-misdirection method (Li et al., 2024), the fine-tuning-against-anchored-rewrites approach of Eldan & Russinovich (2023), and NPO (Zhang et al., 2024), which was introduced to mitigate the catastrophic-collapse failure mode that GA exhibits at large K. The TOFU task (Maini et al., 2024) popularized the synthetic-bio forget-set design we adopt here; the author-clustered bootstrap is our own.

Robustness of LLM unlearning evaluation. Lynch et al. (2024) survey eight evaluation styles for robust LLM unlearning, spanning output, fine-tuning, and latent-knowledge probes. Patil et al. (2024) show that supposedly deleted information is often still recoverable via prompt-based extraction attacks. Our contribution, on reasoning models specifically, is that an additional confound – the trained τ template – makes the naive thinking-leak metric non-identifying as a memorization-residue signal, and that a cheap inference-time template intervention diagnoses this fragility.

Chain-of-thought faithfulness. Chain-of-thought prompting was introduced by Wei et al. (2022), and the question of whether the produced trace actually drives the answer was studied by Turpin et al. (2023) and Lanham et al. (2023). Our finding is consistent with that literature: in the unlearning setting, the trained τ does not faithfully reflect what the unlearned weights still encode – it reflects the upstream training template.