Probe Choice Changes Canary-Memorization Verdicts: Three Post-Hoc Disagreement Case Studies in a Text-Dominant LoRA-Tuned Autoregressive Testbed

Anonymous Authors[1]

Abstract

We audit a fixed prefix-window mean-NLL memorization probe ($K = 20$) on a Qwen2.5-VL-7B canary testbed and report three post-hoc cases where it disagrees with full-span secret NLL or greedy exact-recall. **C3 (false negative, window truncation):** damage lands on hex tokens *outside* $K = 20$; the probe stays flat while hit@1 drops. **C4 (false positive, non-secret drift):** the probe moves, but $\sim 99\%$ sits on non-secret preamble; the secret span and hit@1 are unchanged. **C5 (ambiguous in-window drop):** the probe falls on an undertrained baseline while full-span hex is positive and hit@1 $= 0$. Recommendation: report (i) full-span secret NLL, (ii) a span-localised decomposition, (iii) behavioural exact-recall at $k \geq 4$, and (iv) decoy probes before asserting secret-specificity. Evidence is on controlled canaries in one backbone; magnitudes are testbed-specific.

1. Introduction

A foundational question for deep generative models is whether observed behaviour reflects *generalization* from training data or *memorization* of specific training points (Carlini et al., 2019; 2021; 2023b; Lee et al., 2022; Feldman, 2020). Operationalising this question requires probes that map a model's distribution over outputs to a numeric memorization signal. In the autoregressive setting — which covers language models and the language tower of vision-language and multimodal generative models — three probe families dominate: (a) per-token *negative log-likelihood* (NLL) averaged over a fixed token window (often the first K tokens of a target), (b) *full-span* secret NLL on the substring whose memorization is claimed, and (c) *behavioral exact-recall*, i.e. whether greedy or sampled decoding emits the secret string. Closely related quantities (loss thresholding, perplexity ratios against a held-out reference, calibrated likelihood ratios) appear in adjacent memorization-evaluation literature.

Why probe choice matters. A central concern motivating this paper's audit is that probe choice can change the false-positive and false-negative rates of a memorization claim — a probe-versus-probe consistency question that is separable from the underlying memorization rate. Claims about whether a generative model has memorized rest on which probe was used to call the verdict, and audits asking "does the standard probe agree with itself across natural alternatives?" are a template-specific probe-audit question — one that can confound broader memorization-vs-generalization conclusions if a single window is used uncritically. Liu et al. (2026); Borkar et al. (2025) observe that benign or task-style fine-tuning, or privacy-motivated addition or removal of training records, on data containing *no copy* of a memorized target can move standard memorization probes ("Whack-a-Mole" / privacy-ripple effects); we do not refute or replicate those reactivation findings, but use the same probe family on a different testbed and document where it is and is not self-consistent.

Our testbed. We use a single autoregressive generative-model testbed: a Qwen2.5-VL-7B (Bai et al., 2025) backbone with frozen vision tower and language-tower LoRA (Hu et al., 2022) into which we inject 20 image canaries and 20 text canaries (text canaries pair with a fixed 224×224 mid-gray placeholder image, so the strongest evidence in this paper is essentially text-only). On top of the canary-injected base we stack a second LoRA performing benign supervised fine-tuning (bSFT) on one of four data sources, none of which contain the canary strings (§2). This setup gives us 34 aggregate bSFT cells $=$ 102 seed-level runs (144 with 42 matched-norm LoRA-NOISE controls), and an explicit *undertrained-canary* headroom regime that lets us bound how far reactivation could move the probe in principle (App. A). Across 102 bSFT seed-level runs we observe *no* full-secret NLL recovery and *no* hit@1 above baseline (Tab. 2, App. F); many cells are saturated at hit@1 $= 1.00$ where reactivation cannot be observed without prior secret

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

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

damage, so this null is the *condition under which the probe-disagreement audit is run*, not a substantive null on benign-SFT reactivation in autoregressive generative models. Our setting is adjacent to but distinct from large-scale extraction (Carlini et al., 2021), dedup-driven memorization (Lee et al., 2022), and training-data unlearning (Bourtoule et al., 2021; Maini et al., 2024).

Terminology. We use the following labels throughout:

- *full-secret recovery*: sign-consistent negative $\Delta\mathrm{NLL}_{\mathrm{hex}}$ on the full 13-token secret span across seeds, or hit@1 above baseline.
- *undertrained in-window hex-token NLL improvement*: sign-consistent negative $\Delta\mathrm{NLL}_{\mathrm{mean20}}$ whose per-span decomposition localises onto ~ 9 in-window hex tokens (positions 11–19) plus the leading `canary_lit`, with full-span hex ≥ 0 and hit@1 unchanged.
- *behavioral damage*: secret-span ΔNLL increase *plus* hit@1 drop.
- *non-secret preamble drift*: mean_{20} moves but $\geq 90\%$ of the per-span contribution sits on `preamble`, $\Delta\mathrm{NLL}_{\mathrm{hex}} < 0.5\,|\Delta\mathrm{NLL}_{\mathrm{mean20}}|$, and hit@1 unchanged.
- *window-truncation failure*: mean_{20} flat while damage is visible to wider-window probes or hit@1.

Three probe-disagreement case studies. During the same experiments we exhibit three *post-hoc* cases (Fig. 1) where the internally pre-specified probe $\Delta\mathrm{NLL}_{\mathrm{mean20}}$ gives an answer that differs from full-span secret NLL or from behavioral exact-recall, and each case maps to a distinct failure mode of truncated mean-NLL as an evaluation metric for memorization in an autoregressive generative model:

- **Case 1 (C3, T-bSFT-GUI 5k — false negative under window truncation):** mean_{20} flat ($+0.0001$) while $\Delta_{\mathrm{hex}} = +0.0133$ and hit@1 drops to 0.88 — damage on a hex-token *outside* $K = 20$.
- **Case 2 (C4, T-bSFT-Safety 5k — false positive from non-secret span drift):** $\mathrm{mean}_{20} = +0.0150$ sits $\sim 99\%$ on non-secret `preamble`; secret span and hit@1 unchanged.
- **Case 3 (C5, U-bSFT-GUI 3k — ambiguous undertrained in-window NLL drop):** $\mathrm{mean}_{20} = -0.0070$ on an undertrained baseline localises to in-window hex tokens, but the full-span hex point estimate is positive (and positive across all 3 seeds; hb brackets zero) with hit@1 $= 0$. *No secret-specificity claim* is made; format-prior smoothing or template adaptation are equally consistent without true-vs-decoy probes (a paired decoy audit at the 7B C5 cell remains future work; the smaller-family C5-equivalent decoy results in Tab. 12 are descriptive stress tests at *different* cells).

The $K = 20$ probe and the "no full-secret recovery" operating condition were chosen before seed-level results were aggregated, but were *not* preregistered or externally certified; the three case studies and the selection protocol of §3 are post-hoc.

Contributions.

- A template-specific audit of our internally chosen $K = 20$ truncated mean-NLL probe against full-span secret NLL and behavioral exact-recall, framed as an evaluation-metric question for canary-based memorization in autoregressive generative models (§2).
- Three post-hoc probe-disagreement case studies illustrating two mechanical probe mismatches and one unresolved undertrained-control ambiguity (§3), with full per-cell selection audit (Tab. 4, App. B) and an algebraic span decomposition that closes the dissociation to within bf16 residual (§3.3, App. D).
- Recommendations for evaluation protocols that can flag the two failure modes and the ambiguity case before they propagate into generalization-versus-memorization conclusions (§5, App. A).

Scope. Three case studies in one backbone are not a population claim. We discuss in §4 and §5 which parts of the failure-mode taxonomy are likely to transfer to other autoregressive generative models and which depend on testbed-specific choices (template, tokenizer, LoRA configuration, frozen vision tower). $n = 3$ outer seeds; intervals descriptive only; cross-architecture and cross-scale replication is the natural next experiment.

2. Method: canaries, bSFT matrix, three probes

We instantiate a controlled memorization-evaluation testbed (Fig. 2) whose three ingredients — canary design, benign supervised fine-tuning (bSFT) matrix, and probe specification — jointly determine the generalization-versus-memorization verdict any given probe will return on this autoregressive generative model. The choices below preserve the canary-and-probe setup of Carlini et al. (2019) while making the spans and windows on which probes are computed explicit, so that probe disagreement can be localised to specific token positions rather than absorbed into a global average.

Canaries. Following the small-canary-set tradition of Carlini et al. (2019), $N = 20$ canaries per modality (the exact set size is our own design parameter):

- *Image canaries*: first 20 distinct ScreenSpot v2 screenshots (Wu et al., 2025); target `CANARY-{16hex}-END` where the hex is derived

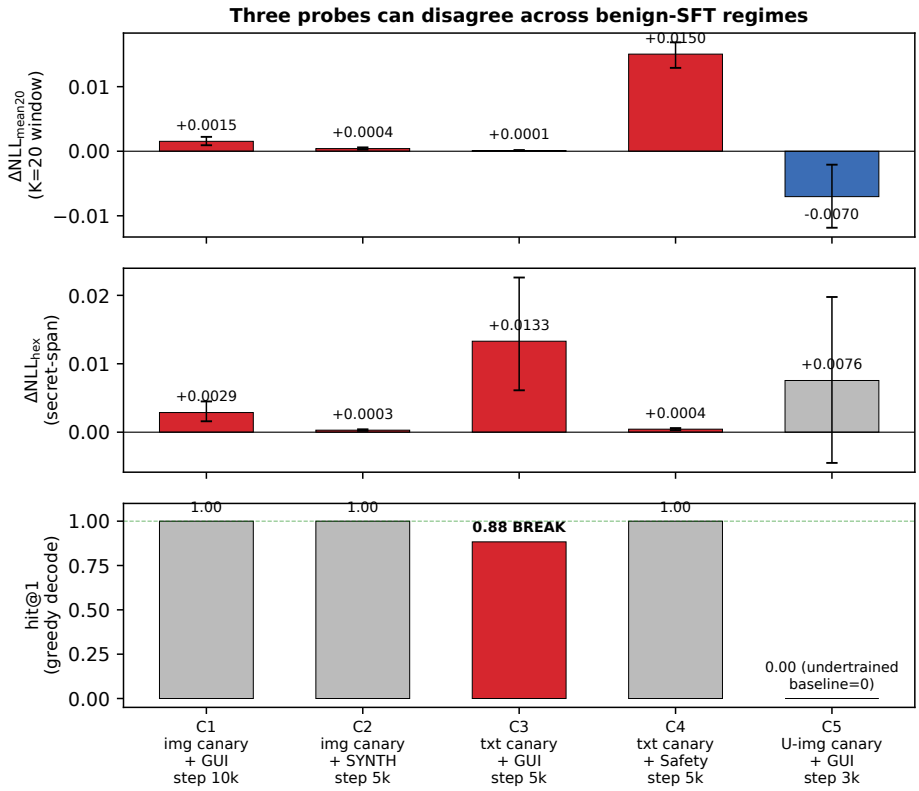

Figure 1. **Three probes can disagree across benign-SFT regimes** (Qwen2.5-VL-7B, stacked LoRA bSFT). C1: img+GUI 10k; C2: img+SYNTH 5k; C3: txt+GUI 5k; C4: txt+SAFETY 5k; C5: U-img+GUI 3k. Top: Δ_{mean20} (internally pre-specified probe); middle: Δ_{hex} (full 13-token secret span); bottom: greedy hit@1. Error bars are descriptive hierarchical resampling intervals (outer 3 seeds, inner 20 canaries, $B = 10{,}000$); they are not population CIs and are not selection-adjusted. **Grayscale-readable encoding:** bar *direction* (sign) carries the qualitative finding; bar *length* carries magnitude; bracket overlap with the dashed 0 baseline indicates uncertainty. Colour (red CI> 0, blue CI< 0, grey brackets 0) is redundant. Only C3 has secret-span NLL and behavioural exact-recall both moving on narrow tail-token damage; mean_{20} misses it (window truncation).

from `sha256("whack-canary-{i:03d}")`.

- *Text canaries*: per-canary user prompts "Provide the diagnostic label for sample id {8hex}" paired with a shared 224×224 mid-gray placeholder image; target `TEXTCAN-{16hex}-END`.

Manifests are SHA-256-pinned on first run, verified each subsequent run. Image-canary screenshots are excluded from bSFT data by filename (perceptual-hash duplicate check in App. J).

Three injection regimes. On Qwen2.5-VL-7B-Instruct (Bai et al., 2025) with frozen vision tower we inject canaries via LoRA (Hu et al., 2022) on the language tower (target modules $\{q, k, v, o, \text{gate}, \text{up}, \text{down}\}$_proj), then `merge_and_unload` the canary LoRA into the base *before* bSFT ("stacked LoRA bSFT" = canary-merged base + bSFT-LoRA loaded at probe time; App. G). All canary LoRAs use $r = 32$, $\alpha = 64$:

- M_{canary} (image, saturated): 80 ep, recall $20/20$.
- $M_{\text{text canary}}$ (text, saturated): 40 ep, recall $20/20$.
- $M_{\text{canary under}}$ (image, undertrained): 10 ep, baseline mean NLL 1.28, hex 2.62, hit $0/20$ (headroom for reactivation).

Non-canary-string SFT (bSFT) variants. "Benign" here denotes the operational property that the SFT data contains no copy of any canary string; SAFETY and RANDOM share the gray-placeholder image with text canaries by construction, only the canary strings are excluded. On each canary-injected base we stack a second LoRA ($r = 16$, $\alpha = 32$, lr 2×10^{-5}, batch 2, grad-accum 4, cosine, 5000 examples) on one of four data sources (none contain canary strings):

- GUI: ScreenSpot grounding (image+text).
- SYNTH: PIL-generated abstract images with templated captions (image+text out-of-domain control).
- SAFETY: `(harmful-request, refusal)` text pairs with placeholder image.
- RANDOM: `(factoid-question,`

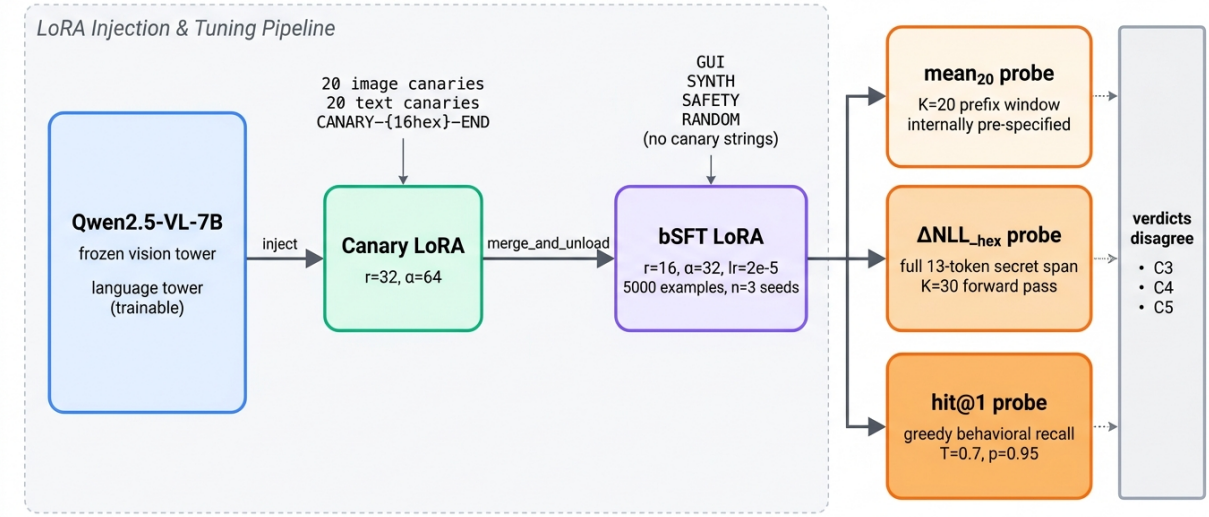

Figure 2. Single-backbone canary-memorization audit testbed. A frozen-vision-tower Qwen2.5-VL-7B receives a canary LoRA (r=32, α=64) carrying 20 image canaries and 20 text canaries with target string `CANARY-{16hex}-END`; the canary LoRA is then `merge_and_unload`-ed into the base before a benign-SFT (bSFT) LoRA (r=16, α=32, lr=2×10^{-5}, 5000 examples, n=3 seeds) is stacked on one of four canary-string-free data sources (GUI, SYNTH, SAFETY, RANDOM). The resulting model is read by three independent memorization probes whose verdicts can disagree across the C3, C4, and C5 case studies: an internally pre-specified mean_{20} prefix-window NLL probe, a full 13-token secret-span $\Delta\mathrm{NLL}_{\mathrm{hex}}$ probe at K=30, and a behavioural greedy-decoding hit@1 probe at T=0.7, p=0.95.

`short-answer`) text pairs with placeholder image.

Matched-norm LoRA-NOISE controls cover image, image-SYNTH, T-GUI, T-Safety, U-image (14 aggregate cells). Step grid (asymmetric, see Tabs. 7–8): GUI-saturated-image at $\{1, 3, 5, 7, 10\}$k; other saturated at $\{1, 3, 5\}$k; undertrained-image at $\{1, 3\}$k; 3 seeds per cell.

All training and probes use bf16; App. D reports the bf16 residual between $K = 20$ probe-pass and $K = 30$ token-level pass for completeness.

Reproducibility fields. Optimizer AdamW, peak lr 2×10^{-5}, cosine schedule, batch 2 with grad-accum 4 (effective 8), max sequence length 1024, canary LoRA epochs as listed above, bSFT seed-level training to 5000 examples (or the cell-specific step grid). RNG seeds $\{0, 1, 2\}$ apply jointly to bSFT data sampling, canary order, optimizer and dropout. Decoding for hit@k uses $T = 0.7$, $p = 0.95$, max new tokens = canary length + 4, and the standard Qwen2.5/Llama-3 chat template with the identical user prompt above; bSFT data is sampled without replacement from each source. Cross-family runs use the same text-only canary subset and the same hyperparameters, modulo removing image fields from the canary template and the bSFT batches.

Tokenization and probe windows. The text-canary assistant target tokenizes deterministically (Table 1); the template extends past token 20. Two consequences: (a) our internally pre-specified $K = 20$ mean-NLL window covers all 10 preamble tokens, the literal, but only 9 of 13 hex tokens, omitting the last 4 hex tokens and the suffix; (b) the secret-span (`canary_hex`) probe widens the forward pass to $K = 30$ to include the full 13 hex tokens.

Notation summary. mean_K always denotes a prefix-window average over positions $0, \ldots, K-1$; mean_{20} is the internally pre-specified prefix probe. hex denotes the fixed 13-token secret span (positions 11–23), evaluated with a $K = 30$ forward pass only to expose the tail tokens; it is not a 30-token average. A $K = 30$ pass is used only to obtain token losses beyond position 19; unless explicitly written as mean_{30}, all hex values are averages over the 13 secret tokens only. All Δ quantities are bSFT-vs.-canary-baseline deltas in NLL loss units (additive, not ratios), aggregated as the per-canary mean averaged across $n = 3$ outer seeds.

Span	Token positions	# tokens	In $K=20$?
`preamble`	0–9	10	all 10
`canary_lit`	10	1	yes
`canary_hex`	11–23	13	first 9 only
`canary_end`	24	1	no (truncated)
`trailing`	25+	–	no (truncated)

Table 1. Token-position layout of the *text*-canary target (Qwen2.5 tokenizer, zero-indexed). The image-canary target `...labeled CANARY-{16hex}-END containing` N `colored shapes...` uses the same span definitions; `canary_lit` is the single token `ARY-` when present (in $\sim 70\%$ of image canaries the leading `CAN` is BPE-absorbed into preamble; in the other $\sim 30\%$ BPE merges `CAN` with `ARY-` so `canary_lit` is empty inside $K=20$, giving the ~ 0.70 average count used in Appendix D). In both templates the last hex tokens and the suffix fall outside $K=20$; per-canary position layouts are deterministic given the SHA-derived hex string.

Three probes. Token positions zero-indexed throughout (Tab. 1); $K = 20 =$ positions 0–19. Let $\ell_\theta(t_j) = -\log p_\theta(t_j \mid t_{<j}, x)$ and $\ell_\theta^{\text{base}}$ the per-token cross-entropy under bSFT and canary-injected baseline. $\Delta\text{NLL}_{\text{mean20}} = \frac{1}{20}\sum_{j=0}^{19}[\ell_\theta(t_j) - \ell_\theta^{\text{base}}(t_j)]$; the truncated-window mean-NLL probe and the $K=20$ cut are our own design choice. $\Delta\text{NLL}_{\text{hex}}$ averages the 13 `canary_hex` tokens (positions 11–23; 11–19 inside $K=20$) with a $K=30$ forward pass; Δ on other spans defined analogously. Throughout the paper, mean_K denotes a prefix-window average over positions $0-(K-1)$, while hex denotes the fixed 13-token secret-span average, even when evaluated with a $K=30$ forward pass for tail-token coverage. hit@k is the fraction of canaries whose literal target appears in $k \in \{1, 4, 16\}$ decoded outputs: greedy at $k=1$ and greedy plus $k-1$ stochastic samples at $T=0.7, p=0.95$ for $k \in \{4, 16\}$ (i.e. 3 and 15 samples beyond greedy).

Aggregation and uncertainty. Each cell pools 20 canaries $\times$ 3 seeds. Main-text statistics are reported as the per-canary mean delta with a 95% *hierarchical bootstrap* CI (outer-resample 3 seeds with replacement, inner-resample 20 canaries within each chosen seed, $B = 10{,}000$, fixed RNG seed). The hierarchical bootstrap accounts for the fact that the same 20 canaries are reused across seeds, unlike a naive IID bootstrap over 60 measurements. With only 3 seeds these CIs are still wide; we report seed ranges in addition wherever a claim is tight, and we do not treat narrow CIs as population-level guarantees. Exact bit-for-bit reproduction is not possible from this submission alone because canary manifests and bSFT split files are not released; the printed tables are intended for numeric audit, not re-running, and the appendix recipe in App. A documents the construction without releasing the artifacts.

Cond.	$\Delta\text{NLL}_{\text{mean20}}$	$\Delta\text{NLL}_{\text{hex}}$	hit@1
C1 img-GUI 10k	$+0.0015$	$+0.0029$	1.00
C2 img-SYNTH 5k	$+0.0004$	$+0.0003$	1.00
C3 txt-GUI 5k	$+0.0001$	$\mathbf{+0.0133}$	**0.88**
C4 txt-SAFETY 5k	$+0.0150$	$+0.0004$	1.00
C5 U-GUI 3k	-0.0070	$+0.0076$	0.00

Table 2. **Probe dissociation across five canonical cells (Δ point estimates).** Qwen2.5-VL-7B, 20 canaries $\times$ 3 seeds. Canary-injected baselines (absolute NLL, not deltas): img= 0.0002/0.0002, txt= 0.0001/0.0003, U-img= 1.2824/2.6184 (mean_{20}/hex). Hierarchical-bootstrap and Student-t_2 95% intervals (descriptive at $n=3$) in App. C and App. F. **C3** (bold): unique saturated cell with hit@1 <1 (Case 1; mean_{20} misses it). C4: drift $\sim 99\%$ on `preamble` (Case 2). C5: in-K=20 hex -0.00583 ($\sim 83\%$) + `canary_lit` -0.00091 ($\sim 13\%$) + bf16 residual -0.00030 = canonical -0.00703 (Case 3; App. D). Full-span hex on C5 is positive in point and across all 3 seeds; descriptive intervals (t_2 excludes zero, hb brackets zero) reported in App. C.

3. Three probe-disagreement case studies

We surface three *post-hoc* case studies (§3.1–§3.3, C3/C4/C5). The matrix is 4 bSFT variants $\times$ 3 canary regimes (saturated image, text, undertrained image) $\times$ 3 seeds $\times$ 3–5 steps, yielding 34 aggregate bSFT cells = 102 seed-level runs (144 with 42 LoRA-NOISE controls). Table 2 shows five canonical cells; full all-cells point estimates in App. F; headline-cell hb and t_2 intervals are listed in App. C.

Selection audit. Case studies are post-hoc (full audit: Tab. 4/App. B). Headline counts:

- **C3** sole hit@1 <1 saturated cell (26-cell pool).
- **C5** largest post-hoc negative mean_{20} among 8 undertrained-image bSFT cells (3/3 seeds same sign).
- **C4** largest of 3 qualifying preamble drift cells.
- Across 48 CI rows, 0 cells have hb-CI screened negative on Δ_{hex}.

Thresholds were chosen before per-cell CIs were aggregated, but were *not* preregistered or externally certified (App. A).

Statistical note ($n=3$). Both hierarchical-bootstrap and Student-t_2 intervals are descriptive only; per-seed sign-consistency describes the selected cells and is *not* inferential evidence after post-hoc selection (App. C, App. A).

"Benign" caveat. "Benign" here means operationally "no canary string in bSFT data," not semantically benign. T-SAFETY/C4 share a fixed gray placeholder image with text canaries (read as "shared-placeholder").

3.1. Case 1 — window truncation misses sparse greedy tail-token brittleness (C3).

Numbers. $\Delta_{\text{mean20}} = +0.0001$ flat; $\Delta_{\text{hex}} = +0.0133$ (3/3 seeds positive); hit@1 drops $1.00 \to 0.88$.

Mechanism. Per-token plot (Fig. 4): tokens 0–22 track baseline within 10^{-4} in absolute NLL; token 23 (the final `canary_hex` BPE piece, 1–2 hex characters before `-END`) absolute NLL jumps to ~ 0.119 (per-token mean across 60 instances), three orders of magnitude above the flat $\sim 10^{-4}$ baseline at neighbouring positions. Position 23 falls outside $K = 20$ (zero-indexed), so mean_{20} is blind to it.

Sparsity. $7/60$ greedy failures on $3/20$ canaries (canary 0 dominates: $+0.122$ alone, leave-one-out hex $+0.0075$); $17/20$ canaries perfect across all seeds; hit@16 $= 58/60$ (App. I). One sparse tail-token counterexample, not broad damage and not distribution-level loss of the secret.

3.2. Case 2 — non-secret preamble drift (C4).

Numbers. $\Delta_{\text{mean20}} = +0.0150$ (3/3 seeds positive); $\Delta_{\text{hex}} = +0.0004$; hit@1 $= 1.00$.

Primary diagnostic (span decomposition). Per-span decomposition (App. D) attributes $\sim 99\%$ of the mean_{20} movement to non-secret `preamble` ($+0.01486/+0.01503$ token-level); only $+0.00017$ falls on `canary_hex`. A probe whose mass sits on non-secret preamble is not secret-span damage.

Appendix-only sensitivity check (merge-and-unload). Merging bSFT-LoRA into the base collapses Δ_{preamble} from $+0.0297$ to $+0.0014$ (-95%); C1/C3/C5 secret-span ratios stay $0.98/0.98/0.77$ and C4's own `canary_hex` survives at 0.51 (Tab. 10). The collapse is C4-preamble-specific $\Rightarrow$ evaluation-stack-dependence on preamble (no causal claim). This ablation is diagnostic of evaluation-stack sensitivity only; the primary evidence for Case 2 is the per-span decomposition above.

3.3. Case 3 — unresolved undertrained-control ambiguity (C5).

Numbers. $\Delta_{\text{mean20}} = -0.0070$ (3/3 seeds negative); full-span $\Delta_{\text{hex}} = +0.0076$; hit@1 $= 0$.

Decomposition (App. D): in-window hex -0.00583 + `canary_lit` -0.00091 + bf16 residual -0.00030.

Reading. An in-window hex-token NLL drop on an undertrained baseline; *no* secret-specificity claim. Format-prior smoothing or template adaptation are equally consistent without true-vs-decoy probes.

Per-seed sign consistency for headline cells.

Cell	Probe	seed 0	seed 1	seed 2	Sign
C3	Δ_{hex}	$+0.0151$	$+0.0146$	$+0.0102$	3/3+
C4	Δ_{mean20}	$+0.0130$	$+0.0161$	$+0.0161$	3/3+
C5	Δ_{mean20}	-0.0063	-0.0082	-0.0066	3/3–

C5 algebra closure.

$$\underbrace{-0.00583}_{\text{in-K hex}} + \underbrace{-0.00091}_{\text{canary\_lit}} + \underbrace{+0.00001}_{\text{preamble}} + \underbrace{-0.00030}_{\text{bf16}} = -0.00703$$

3.4. Window K-sweep: 7B failure modes vs. smaller-family stress test

For the smaller-family stress test (§3.5) we re-aggregate the per-token NLL files at $K \in \{10, 15, 20, 25, 30\}$; for the 7B headline cells we report only the derivable extremes $K = 10$, $K = 20$, and the full `canary_hex` span (the intermediate 7B $K \in \{15, 25, 30\}$ values would require re-running the per-token aggregator with a different window length and are not printed in this submission).

On the 7B headline cells, the algebra closure (Tab. 5) plus the full-span readout (Tab. 2) pin down the printed extremes (see Tab. 6). They imply: **C3** is $\sim +10^{-4}$ for $K \leq 20$ and only diverges from the in-window value once the tail-token spike at position 23 enters the window; **C4** stays positive across $K = 10 \to 20$, with the preamble-only $K = 10$ mean ($+0.0297$) accounting for $\sim 99\%$ of the canonical $K = 20$ value after rescaling by $10/20$ (i.e. $0.5 \times 0.0297 = 0.01486$ vs. $+0.01504$); **C5**'s mean is negative at the canonical $K = 20$ but turns positive at the full-span readout, the joint-product signature of window truncation *and* the undertrained-baseline regime.

On the smaller-family C3/C4/C5-equivalent cells (Fig. 3): the K-monotone shape on Qwen2.5-1.5B and Llama-3.2-1B differs from the 7B testbed: in particular, the negative-mean signature behind 7B Case 3 does *not* reproduce on either smaller family at the C5-equivalent cell, and Llama-1B's Δmean_K at C3/C4-eq is dominated by a catastrophic preamble drift at every K rather than by the tail-token / preamble-fraction mechanism we observe on 7B.

3.5. Descriptive smaller-family non-replication check

We run a smaller-family stress test at the C3/C4/C5-equivalent cells on Qwen2.5-1.5B-Instruct (Qwen et al., 2024) and Llama-3.2-1B-Instruct (Meta AI, 2024) (text-only LMs, same canary template, same bSFT data minus image fields, same probe family, $n = 3$ outer seeds at each cell). We do not print or claim a full 4×3 smaller-family bSFT grid; App. F is the 7B all-cells matrix only. Headline-cell point estimates appear in Tab. 3; we read these as descriptive stress tests on smaller text-only backbones, not as case-level replication of the 7B mechanisms (Tab. 3 caption).

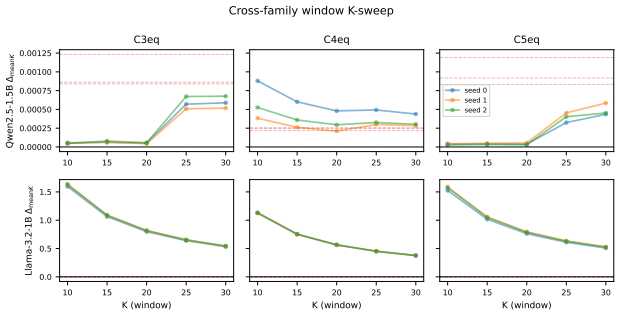

Figure 3. **Descriptive smaller-family stress test:** $\Delta\mathrm{NLL}_{\mathrm{mean}K}$ for $K \in \{10, 15, 20, 25, 30\}$ on Qwen2.5-1.5B (top) and Llama-3.2-1B (bottom) at the C3/C4/C5-equivalent cells, per seed (descriptive at $n = 3$). *Y-axis scales differ by family by* $\sim 1000\times$ (Llama deltas are at 10^0 scale, Qwen at 10^{-3}); we read this figure as a stress test on smaller text-only LMs, not as case-level replication. Dashed red line marks the full `canary_hex` span Δ where present. Qwen-1.5B is near-flat through $K = 20$ at C3-eq while the full-hex Δ is positive; Llama-1B instead shows large preamble-dominated deltas across K at C3/C4-eq; C5-eq has no negative-mean signature on either family (Case 3's mechanism does not have a direct analog in this protocol).

Table 3. **Smaller-family stress test on the C3/C4/C5-equivalent cells.** Point-estimate ΔNLL (mean_{20} and full canary_hex span) and greedy hit@1 at $n = 3$ outer seeds. The 7B row reproduces Tab. 2; the 1.5B / 1B rows are descriptive stress-test runs on smaller text-only LMs at cells fixed by the 7B labelling, not case-level replication of the 7B mechanisms. Qwen-1.5B C3/C4/C5-eq are near-zero with greedy hit@1 $= 1.00$ throughout, and Llama-1B exhibits a different catastrophic-mean / preserved-hex / broken-greedy disagreement pattern at every cell.

Family	Cell	$\Delta_{\mathrm{mean}20}$	Δ_{hex}	hit@1
Qwen2.5-VL-7B	C3-eq	$+0.0001$	$+0.0133$	0.88
Qwen2.5-1.5B	C3-eq	$+0.0001$	$+0.0010$	1.00
Llama-3.2-1B	C3-eq	$+0.8104$	$+0.0006$	0.37
Qwen2.5-VL-7B	C4-eq	$+0.0150$	$+0.0004$	1.00
Qwen2.5-1.5B	C4-eq	$+0.0003$	$+0.0002$	1.00
Llama-3.2-1B	C4-eq	$+0.5643$	$+0.0001$	0.00
Qwen2.5-VL-7B	C5-eq	-0.0070	$+0.0076$	0.00
Qwen2.5-1.5B	C5-eq	$+0.0000$	$+0.0010$	1.00
Llama-3.2-1B	C5-eq	$+0.7814$	$+0.0006$	0.00

Cross-family verdict. The smaller-family stress test surfaces *additional* probe disagreements but does *not* replicate all three 7B case mechanisms at the C3/C4/C5-equivalent cells. Case 3's undertrained-baseline mechanism has no direct analog in our text-only cross-family setup (text LMs in this protocol have no separate undertrained-canary regime; the image-canary gray-placeholder construction that triggers C5 on the VLM testbed is unavailable on text-only LMs), and the C5 cell at the 7B-fixed location does not reproduce as a misleading negative on either family (Tab. 3: Qwen-1.5B C5-eq $\Delta_{\mathrm{mean}20} = +0$; Llama-1B C5-eq $+0.78$ matches the same catastrophic-mean / preserved-hex / broken-greedy pattern as Llama C3-eq and C4-eq, not Case 3). Two additional smaller-family phenomena — a migrated Qwen-1.5B C4-equivalent cell with 98% preamble-driven $\Delta_{\mathrm{mean}20} = +0.22$, and a Llama-3.2-1B T-bSFT-GUI step-trajectory where $\mathrm{mean}_{K=20}$ grows monotonically while greedy hit@16 recovers across step — and the smaller-family decoy audit at the candidate C5-equivalent cells (showing additive NLL gaps of $+1.67$ to $+10.86$ loss units between shuffled-hex / wrong-secret / format-only decoys and the true target) are reported in App. K; this is a calibration of the decoy construction at *different* cells and does not adjudicate the original 7B C5 ambiguity, which lacks a paired decoy audit. Magnitudes are 3-seed point estimates, not inferential, and we read these rows as descriptive smaller-family stress tests rather than as case-level replication.

4. Related Work

Memorization probes for autoregressive models. The canary-and-probe paradigm we use originates with Secret Sharer (Carlini et al., 2019), which introduced *exposure* as a calibrated rank-based metric for unintended memorization, and was generalized in Carlini et al. (2023b) to discoverable-extraction rates across model scales, data duplication, and prompt length. Truncated mean-NLL over the first K target tokens — the probe we audit — is a natural prefix-window operationalization in this lineage; the present paper's three case studies show that, on a single-backbone testbed, the choice of probe (truncated mean NLL, full-span loss, behavioural exact-recall) can materially change false-positive and false-negative rates. We do not propose a new probe; we document, in one autoregressive generative-model testbed, three concrete cases where truncated mean-NLL silently disagrees with full-span NLL or with behavioral exact-recall, and attribute each disagreement to a specific token-position or baseline-regime mechanism.

Extraction and large-scale memorization. Carlini et al. (2021) demonstrated that training data can be recovered from large language models via query-based extraction with ranking-based filtering, shifting the field toward extraction-based memorization metrics that do not depend on a fixed-window NLL probe at all; subsequent work (Nasr et al., 2025) scales these attacks to production language models. Memory-trace studies that frame memorization as a finetuning-time and forgetting-time quantity (Jagielski et al., 2023) sit between extraction-based and NLL-based metrics and motivate the need to know what fixed-window NLL does and does not measure on a controlled canary set. Deduplication-driven studies (Lee et al., 2022) establish that memorization scales with duplication count, providing a population-level link between training distribution and probe response that any single-window probe must respect.

Our setting is adjacent: the canary distribution is deliberately controlled, and we audit the probe rather than the underlying memorization rate, but the same evaluation-metric concerns apply.

Benign or safety-driven side effects on memorization. The closest motivating works are Liu et al. (2026); Borkar et al. (2025), which report that benign or task-style fine-tuning, or privacy-motivated addition or removal of training records, on data containing no copy of a memorized target can move standard memorization probes ("Whack-a-Mole" / privacy-ripple effects). These claims are exactly the kind of memorization-versus-generalization question that probe choice can flip — the same probe-window decision that buries Case 1 (a tail-token greedy miss outside $K = 20$) can also buries any reactivation that occurs predominantly outside the chosen window. We do not claim our three case studies refute or replicate those reactivation findings; we use the same probe family on a different testbed and document where it is and is not self-consistent.

Unlearning and lifecycle. A separate line studies training-data unlearning, both in the general SISA-style sense (Bourtoule et al., 2021) and as targeted forgetting of memorized content in generative models (Maini et al., 2024). The probe-disagreement phenomena we document are a precondition for unlearning evaluation: if truncated mean-NLL drops on an undertrained baseline without secret-specific evidence (Case 3), then unlearning audits using the same probe family will register apparent "forgetting" that does not correspond to a behavioural change. We adapt the decoy/canary control of Secret Sharer (Carlini et al., 2019) to expose this ambiguity in our setting and suggest the same control is useful in unlearning evaluation.

Memorization in non-autoregressive generative models. Memorization in diffusion models is an active area of the foundations literature (Carlini et al., 2023a; Somepalli et al., 2023a;b), and the metric-dependence concern we document is not specific to autoregressive backbones: visual generative models can be probed via near-duplicate retrieval, likelihood-based membership inference, or direct extraction, each of which has its own false-positive and false-negative profile relative to true training-image replication. We do not run experiments on diffusion models in this paper; we limit our claims to an autoregressive setting and discuss in §5 which parts of the failure-mode taxonomy plausibly carry over and which require dedicated experiments.

What this paper adds. Relative to the works above, our contribution is narrow and audit-shaped: we treat the truncated mean-NLL probe as the unit of analysis, decompose it onto explicit token spans (`preamble`, `canary_lit`, `canary_hex`, `canary_end`, `trailing`; Tab. 1) in one testbed, and report three concrete cases where the span-decomposition or behavioural-recall verdict diverges from the truncated probe. Three cases in one backbone are not a population claim; the contribution is to make the failure-mode taxonomy *nameable* so that probe-window choices in future memorization-versus-generalization studies can be made and reported with the relevant audits attached.

5. Discussion: evaluation protocols for memorization in generative models

Take-home. In this testbed, truncated $\mathrm{mean}_{K=20}$ NLL shows two well-localised disagreement modes (Case 1 false negative, Case 2 false positive) and one ambiguity mode (Case 3: an undertrained in-window NLL drop without full-secret recovery, motivating decoy controls; we do not claim secret-specific recovery on Case 3). Memorization audits on canary sets should therefore report, beyond $\Delta\mathrm{NLL}_{\mathrm{mean20}}$ alone:

- **(a) Full-span secret NLL.** A window that omits any portion of the secret span is blind to damage on the omitted positions. Case 1: in our tokenizer the last hex BPE piece falls at position 23, outside $K = 20$, so a $0-19$ window silently drops it.
- **(b) Behavioural hit@k with $k \geq 4$.** Greedy hit@1 catches Case 1's tail-token brittleness when likelihood-based probes do not; sampled hit@k bounds whether the secret persists in the distribution. In our matrix hit@16 $=$ $58/60$ on C3 — broken canaries remain in the secret distribution even when greedy decoding breaks on a single tail token.
- **(c) Per-span decomposition (e.g. `preamble` vs. secret).** Isolates whether a probe move is on the secret or on template-shared positions. Case 2's $+0.0150$ probe move sits $\sim 99\%$ on the non-secret `preamble` span while the secret-span Δ is $+0.0004$ and hit@1 $= 1.00$.
- **(d) A saturated and an undertrained control** to bound how far reactivation could move the probe in principle. Across the 4×2 undertrained-image bSFT grid plus matched-norm LoRA-NOISE, no variant produces verbatim hit@1 > 0 (App. A), so probe drops in this regime cannot be attributed to secret recovery without further evidence.
- **(e) Decoy probes (shuffled / wrong-secret / format-only)** before asserting secret-specificity. Case 3: an in-window hex-token NLL drop is consistent with both secret-token learning and format-prior smoothing. A 7B C5 decoy remains future work (cf. App. K).

References

Bai, S., Chen, K., Liu, X., et al. Qwen2.5-VL technical report, 2025.

Borkar, J., Jagielski, M., Lee, K., Mireshghallah, N., Smith, D. A., and Choquette-Choo, C. A. 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. Association for Computational Linguistics, 2025.

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)*, pp. 141–159. IEEE, 2021.

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

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

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, pp. 5253–5270, Anaheim, CA, August 2023a. USENIX Association.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023b.

Cheng, K., Sun, Q., Chu, Y., Xu, F., Li, Y., Zhang, J., and Wu, Z. SeeClick: Harnessing GUI grounding for advanced visual GUI agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9313–9332, 2024.

Feldman, V. Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing (STOC)*, pp. 954–959, 2020.

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 *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

Jagielski, M., Thakkar, O., Tramèr, F., Ippolito, D., Lee, K., Carlini, N., Wallace, E., Song, S., Guha Thakurta, A., Papernot, N., and Zhang, C. Measuring forgetting of memorized training examples. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8424–8445. Association for Computational Linguistics, 2022.

Liu, X., Mireshghallah, N., Ginsburg, J. C., and Chakrabarty, T. Alignment Whack-a-Mole: Finetuning activates verbatim recall of copyrighted books in large language models, 2026. arXiv preprint arXiv:2603.20957, under review.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling (COLM)*, 2024.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices 2024. Released 2024-09-25; accessed 2026-05-09.

Nasr, M., Rando, J., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from aligned, production language models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Qwen et al. Qwen2.5 technical report, 2024. arXiv preprint arXiv:2412.15115.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, June 2023a.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 47783–47803, 2023b.

Wu, Z., Wu, Z., Xu, F., Wang, Y., Sun, Q., Jia, C., Cheng, K., Ding, Z., Chen, L., Liang, P. P., and Qiao, Y. OS-ATLAS: A foundation action model for generalist GUI agents. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

A. Scope of claims and reproducibility

Scope of claims. The three case studies hold in our testbed only: Qwen2.5-VL-7B-Instruct, language-tower LoRA with frozen vision tower, four bSFT variants (GUI, SYNTH, SAFETY, RANDOM), 20 canaries per modality (image, text, undertrained-image), exact substring probes, $K = 20$ for mean-NLL. Across the 4×2 undertrained-image bSFT grid plus matched-norm LoRA-NOISE (2 steps), no variant produces verbatim hit@1 > 0 on the undertrained canaries. The strongest evidence (Case 1, C3) is essentially text-only; we do not extrapolate to other backbones, PEFT configurations, canary templates, or non-autoregressive families (diffusion, flow). $n = 3$ seeds; hierarchical-bootstrap and Student-t_2 intervals are descriptive only, and after post-hoc selection no inferential claim survives selection adjustment (Tab. 4).

Reproducibility. We do not release supplementary material. All canonical numbers are reproduced in the printed tables: per-cell deltas in Table 8, per-seed sign-consistency and headline intervals in App. C, span decomposition in Table 5, and merge/stacked comparisons in Tables 9–10. Selection-audit thresholds in Table 4 ($|\Delta_{\text{mean20}}| > 10^{-3}$, $|\Delta_{\text{hex}}| < 0.5|\Delta_{\text{mean20}}|$, baseline + post hit@1 $= 1$, screened negative under hb interval) were chosen before per-cell CIs were aggregated but were *not* preregistered or externally certified; read them as *authors' internal selection thresholds*, not a registered analysis plan. The full pipeline is recoverable from the body and tables; exact bit-for-bit reproduction requires the unreleased canary manifests and bSFT splits. Canary secrets are SHA-256 of `whack-canary-{i:03d}` truncated to 16 hex; bSFT samples 5000 examples without replacement under seed $s \in \{0, 1, 2\}$ with image-filename exclusion (§J); training is stacked LoRA ($r = 16$, $\alpha = 32$) on top of the canary-merged base with AdamW, peak lr 2×10^{-5}, cosine schedule, batch $2 \times$ accum 4, max sequence length 1024.

B. Selection audit (full)

Family	Pool	#qual.	Shown
hit@1 < 1 at saturated baseline	26 saturated cells (img+txt)	1	C3 (unique)
preamble drift: $\lvert\Delta_{\text{mean20}}\rvert > 10^{-3}$, $\lvert\Delta_{\text{hex}}\rvert < 0.5\lvert\Delta_{\text{mean20}}\rvert$, baseline + post hit@1 $= 1$	26 saturated cells	3 (T-Safety 1/3/5k)	C4 (largest magnitude)
undertrained-regime negative mean$_{20}$: $\Delta_{\text{mean20}} < -10^{-3}$	8 undertrained-image bSFT cells	2 (U-GUI 1k/3k)	C5 (largest magnitude = U-GUI 3k)
screened negative under hb interval on Δ_{hex}	48-row CI dump	0	—
descriptively *negative* hb interval on Δ_{mean20}	48	1 (U-GUI 3k=C5)	C5 (sole negative)

Table 4. **Selection audit.** Eligible pool, threshold, and qualifier count for every diagnostic family we examined. Thresholds were chosen before per-cell CIs were aggregated, not externally certified (App. A). **C3** is the only hit@1 < 1 saturated cell; **C5** is selected as the largest post-hoc negative mean$_{20}$ point estimate among the 8 undertrained-image bSFT cells; its hb interval is the only negative mean$_{20}$ interval among the 48-row CI pool, reported descriptively only; **C4** is the largest of 3 qualifying preamble-drift cells. **C1, C2** are diagnostic baselines (largest-step saturated GUI; saturated SYNTH 5K, sub-threshold). The full matrix is reported in Table 8; we do not claim outcome-blind selection. $\rho(\Delta_{\text{mean20}}, \Delta_{\text{hex}}) = 0.82$ across 34 bSFT cells.

C. Per-seed sign-consistency and headline intervals

Across the five canonical cells, every per-seed $\Delta\text{NLL}_{\text{mean20}}$, $\Delta\text{NLL}_{\text{hex}}$, and hit@1 value is sign-consistent across the 3 outer seeds (max spread 0.0049 on C3 hex, 0.0255 on C5 `canary_lit`); the 3-seed mean of each cell equals Table 2 by construction. Headline-row Student-t_2 vs. hier-bootstrap intervals: **C3 hex** $[+0.0066, +0.0200]$ / $[+0.0061, +0.0226]$; **C4 mean$_{20}$** $[+0.0106, +0.0195]$ / $[+0.0129, +0.0169]$; **C5 mean$_{20}$** $[-0.0096, -0.0045]$ / $[-0.0118, -0.0021]$; **C5 full hex** $[+0.0027, +0.0124]$ / $[-0.0045, +0.0198]$. Both interval families are descriptive at $n = 3$; we read claims as joint requirements (same direction in both) and mark *hb-only* any claim that relies on hier-bootstrap alone.

D. Algebra closure: span decomposition of $\Delta\text{NLL}_{\text{mean20}}$

For each canonical cell we decompose $\Delta\text{NLL}_{\text{mean20}}$ into per-span contributions *inside* the $K = 20$ window. For each (canary, seed) instance let $n_{s,c}$ be the number of tokens of span s that fall inside the canary's $K = 20$ window and $\overline{\Delta_{s,c}}$ the per-token ΔNLL over those tokens; the per-instance contribution of span s is $n_{s,c}\,\overline{\Delta_{s,c}}/20$ (0 if the canary has no tokens of s inside $K = 20$). Averaging across all 60 (canary, seed) instances and summing across spans yields the per-canary mean $\Delta\text{NLL}_{\text{mean20}}$ from the *token-level* forward pass.

The canonical mean$_{20}$ uses the $K = 20$ probe pass; the token-level pass uses $K = 30$ to cover the hex span. Both compute

teacher-forced cross-entropy in bf16 from the same weights, but bf16's 7-bit mantissa produces a per-cell residual of order 10^{-4} (max -0.00030 on C5). The residual is computed independently as canonical minus enumerated and shown as a disclosure row, not a free parameter. For *text* canaries (C3, C4) the Qwen2.5 tokenizer yields exactly $n_{\text{preamble}}=10$, $n_{\text{lit}}=1$, $n_{\text{hex}}=9$ inside $K=20$; for *image* canaries (C1, C2, C5) BPE boundaries vary across strings, with cell-level averages $\overline{n}_{\text{lit}}\approx 0.70$, $\overline{n}_{\text{hex}}\approx 9.30$.

Cell	Span	$\overline{\Delta_s}$ /tok	$\overline{n_s}$	Contrib.
C1 (img)	`canary_hex`	$+0.00317$	9.30	$+0.00148$
	`canary_lit`	-0.00004	0.70	-0.00000
	`preamble`	$+0.00012$	10.00	$+0.00006$
	enumerated sum (token-level $K=30$)		20.00	$+0.00154$
	bf16 precision residual		—	$<10^{-5}$
	canonical (Tab. 2)		20.00	$\mathbf{+0.00154}$
C2 (img)	`canary_hex`	$+0.00041$	9.30	$+0.00019$
	`canary_lit`	$+0.00004$	0.70	$+0.00000$
	`preamble`	$+0.00044$	10.00	$+0.00022$
	enumerated sum (token-level $K=30$)		20.00	$+0.00041$
	bf16 precision residual		—	$<10^{-5}$
	canonical (Tab. 2)		20.00	$\mathbf{+0.00041}$
C3 (text)	`canary_hex`	$+0.00013$	9.00	$+0.00006$
	`canary_lit`	$+0.00000$	1.00	$+0.00000$
	`preamble`	$+0.00006$	10.00	$+0.00003$
	enumerated sum (token-level $K=30$)		20.00	$+0.00009$
	bf16 precision residual		—	$<10^{-5}$
	canonical (Tab. 2)		20.00	$\mathbf{+0.00009}$
C4 (text)	`canary_hex`	$+0.00037$	9.00	$+0.00017$
	`canary_lit`	$+0.00000$	1.00	$+0.00000$
	`preamble`	$+0.02973$	10.00	$+0.01486$
	enumerated sum (token-level $K=30$)		20.00	$+0.01503$
	bf16 precision residual		—	$+0.00001$
	canonical (Tab. 2)		20.00	$\mathbf{+0.01504}$
C5 (U-img)	`canary_hex`	-0.01253	9.30	-0.00583
	`canary_lit`	-0.02601	0.70	-0.00091
	`preamble`	$+0.00002$	10.00	$+0.00001$
	enumerated sum (token-level $K=30$)		20.00	-0.00673
	bf16 precision residual		—	-0.00030
	canonical (Tab. 2)		20.00	$\mathbf{-0.00703}$

Table 5. Exact span decomposition of $\Delta\text{NLL}_{\text{mean20}}$ across the five canonical cells. "(text)" / "(img)" / "(U-img)" marks the canary modality (text canaries deterministically tokenize to $n_{\text{lit}}=1$, $n_{\text{hex}}=9$ inside $K=20$; image canaries are sample-averaged at $\overline{n}\approx 0.70/9.30$). **Contrib.** is the per-instance contribution averaged across all 60 (canary, seed) instances (0 for canaries lacking that span inside $K=20$): $\langle n_{s,c}\,\overline{\Delta_{s,c}}/20\rangle$. **Enumerated sum** reproduces the per-canary mean of the token-level forward pass. **bf16 precision residual** is computed independently as canonical minus enumerated; we display it as a disclosure row, not as a free parameter (max $|\cdot|=0.00030$ on C5). **Canonical** is identically the Table 2 value. For **C5**, the -0.0070 canonical mean20 decomposes into a $\sim$9.3-token in-K=20 hex drop (~ -0.013/tok, contribution -0.00583, $\sim 83\%$ of canonical) and a partial `canary_lit` drop (~ -0.026/tok where present in $\sim 70\%$ of canaries, contribution -0.00091, $\sim 13\%$). The full-span hex point estimate $\Delta\text{NLL}_{\text{hex}}=+0.0076$ in Table 2 stays positive because out-of-K=20 hex tail tokens degrade by $\sim +0.06$/tok on average; hit@1 $=0.00$ across all seeds. Sign-consistency: all 3 seeds give negative mean_{20} and positive full-span hex.

7B window K-sweep — derivable extremes. The K-monotone claims for the 7B headline cells are recoverable from the algebra closure plus the headline full-span values: $K=10$ is the preamble-only mean, $K=20$ is the algebra-closure mean, and the full hex column reproduces Tab. 2. Intermediate $K\in\{15,25,30\}$ values require re-running the per-token aggregator and are not printed.

Cell	$\text{mean}_{K=10}$ (preamble-only)	$\text{mean}_{K=20}$	hex (full 13-tok)
C3 (txt-GUI 5k)	$+0.0001$	$+0.0001$	$+0.0133$
C4 (txt-SAFETY 5k)	$+0.0297$	$+0.0150$	$+0.0004$
C5 (U-GUI 3k)	$+0.0000$	-0.0070	$+0.0076$

Table 6. **7B window extremes for the headline cells (derivable from the algebra closure).** $\text{mean}_{K=10}$ is the per-canary preamble-only mean (10 preamble tokens divided by $K=10$), so it equals the per-token preamble $\overline{\Delta}_{\text{preamble}}$ in Tab. 5; $\text{mean}_{K=20}$ matches Tab. 2; the full `canary_hex` column is the per-canary mean over the 13-token secret span (positions 11–23, $K=30$ pass). **C3** stays at $+10^{-4}$ for $K\le 20$ and only diverges from the in-window value once the tail-token spike at position 23 enters the window; **C4** is positive at $K=10$ with the preamble already absorbing the bulk and decays toward the small full-hex value as more hex positions enter; **C5** is negative at $K=20$ but turns positive at the full-hex readout. Point estimates only ($n=3$ outer seeds, descriptive).

E. Cell accounting and matrix counts

The matrix decomposes as 4 bSFT variants (GUI, SYNTH, SAFETY, RANDOM) $\times$ 3 canary regimes (saturated image, text, undertrained image) $\times$ 3 seeds $\times$ 3–5 step counts.

Bucket	**Aggr. rows**	**Seed-level runs**	**In "34" denom.?**	**CIs computed?**
bSFT (GUI/SYNTH/Safety/Random)	34	102	yes	yes
LoRA-Noise norm-matched controls	14	42	no	yes
Baselines (0-step rows)	3	9	no	no
CI-bearing rows	*48* (=34+14)	—	—	—
Printed all-cells table rows	*51* (=48+3)	—	—	—
Total seed-level runs	—	144 (=102+42)	—	—

Table 7. Explicit accounting. The 34-cell denominator covers only the GUI/SYNTH/Safety/Random bSFT regimes (image:14, text:12, undertrained image:8); the CI-bearing pool adds 14 LoRA-Noise controls (baselines have $\Delta = 0$); printed Table 8 additionally lists 3 baseline rows.

F. All-cells systematic table

Table 8 reports $\Delta\mathrm{NLL}_{\mathrm{mean20}}$, $\Delta\mathrm{NLL}_{\mathrm{hex}}$, and hit@1 for every (canary modality, variant, step) cell in our matrix, averaged across 3 seeds. Per-cell hierarchical-bootstrap ($B = 10000$) and between-seed Student-t_2 95% CIs on $\Delta\mathrm{NLL}_{\mathrm{mean20}}$ and $\Delta\mathrm{NLL}_{\mathrm{hex}}$ are computed for the 48 CI-bearing rows (34 bSFT + 14 LoRA-Noise; baselines have $\Delta = 0$ by construction); summary counts for the headline cells are reported in App. C.

No descriptively clear-of-zero negative hex anywhere. Across all 48 CI-bearing cells, *zero* cells have a hierarchical-bootstrap 95% CI for $\Delta\mathrm{NLL}_{\mathrm{hex}}$ strictly below zero. Three cells have a slightly negative point estimate (U-LORA-NOISE -0.00113, U-BSFT-GUI -7×10^{-5}, U-BSFT-SAFETY -0.00062; all undertrained-regime step 1000, where baselines have hex NLL 2.6 and hit@1 = 0); each brackets zero under both hier-bootstrap and t_2 between-seed intervals.

Mod.	Variant	Step	Δ_{mean20}	Δ_{hex}	hit@1	n_{seeds}
img	LoRA-Noise	1000	$+6e-06$	$+1e-05$	1.00	3
img	LoRA-Noise	3000	$+1e-05$	$+2e-05$	1.00	3
img	LoRA-Noise	5000	$+1e-05$	$+2e-05$	1.00	3
img	LoRA-Noise-SYNTH	1000	$+8e-06$	$+1e-05$	1.00	3
img	LoRA-Noise-SYNTH	3000	$+1e-05$	$+2e-05$	1.00	3
img	LoRA-Noise-SYNTH	5000	$+9e-06$	$+2e-05$	1.00	3
img	bSFT-GUI	1000	$+4e-05$	$+6e-05$	1.00	3
img	bSFT-GUI	3000	$+0.0002$	$+0.0003$	1.00	3
img	bSFT-GUI	5000	$+0.0004$	$+0.0007$	1.00	3
img	bSFT-GUI	7000	$+0.0005$	$+0.0008$	1.00	3
img	bSFT-GUI	10000	$+0.0015$	$+0.0029$	1.00	3
img	bSFT-Random	1000	$+3e-05$	$+5e-05$	1.00	3
img	bSFT-Random	3000	$+9e-06$	$+2e-05$	1.00	3
img	bSFT-Random	5000	$+1e-05$	$+2e-05$	1.00	3
img	bSFT-SYNTH	1000	$+0.0002$	$+0.0001$	1.00	3
img	bSFT-SYNTH	3000	$+0.0003$	$+0.0003$	1.00	3
img	bSFT-SYNTH	5000	$+0.0004$	$+0.0003$	1.00	3
img	bSFT-Safety	1000	$+2e-05$	$+4e-05$	1.00	3
img	bSFT-Safety	3000	$+2e-06$	$+1e-05$	1.00	3
img	bSFT-Safety	5000	$+2e-06$	$+1e-05$	1.00	3
img	baseline	0	$+0e+00$	$+0e+00$	1.00	3
txt	T-LoRA-Noise-GUI	1000	$+2e-06$	$+4e-05$	1.00	3
txt	T-LoRA-Noise-GUI	3000	$+7e-06$	$+3e-05$	1.00	3
txt	T-LoRA-Noise-GUI	5000	$+1e-05$	$+8e-05$	1.00	3
txt	T-LoRA-Noise-Safety	1000	$+3e-06$	$+2e-05$	1.00	3
txt	T-LoRA-Noise-Safety	3000	$+7e-06$	$+4e-05$	1.00	3
txt	T-LoRA-Noise-Safety	5000	$+7e-06$	$+3e-05$	1.00	3
txt	T-bSFT-GUI	1000	$+3e-05$	$+0.0002$	1.00	3
txt	T-bSFT-GUI	3000	$+7e-05$	$+0.0027$	1.00	3
txt	T-bSFT-GUI	5000	$+9e-05$	$+0.0133$	0.88	3
txt	T-bSFT-Random	1000	$+2e-05$	$+0.0002$	1.00	3
txt	T-bSFT-Random	3000	$+2e-05$	$+0.0002$	1.00	3
txt	T-bSFT-Random	5000	$+3e-05$	$+0.0002$	1.00	3
txt	T-bSFT-SYNTH	1000	$+0.0003$	$+0.0007$	1.00	3
txt	T-bSFT-SYNTH	3000	$+0.0011$	$+0.0012$	1.00	3
txt	T-bSFT-SYNTH	5000	$+0.0015$	$+0.0013$	1.00	3
txt	T-bSFT-Safety	1000	$+0.0076$	$+0.0003$	1.00	3
txt	T-bSFT-Safety	3000	$+0.0097$	$+0.0004$	1.00	3
txt	T-bSFT-Safety	5000	$+0.0150$	$+0.0004$	1.00	3
txt	T-baseline	0	$+0e+00$	$+0e+00$	1.00	3
U-img	U-LoRA-Noise	1000	-0.0003	-0.0011	0.00	3
U-img	U-LoRA-Noise	3000	-0.0001	$+0.0011$	0.00	3
U-img	U-bSFT-GUI	1000	-0.0033	$-7e-05$	0.00	3
U-img	U-bSFT-GUI	3000	-0.0070	$+0.0076$	0.00	3
U-img	U-bSFT-Random	1000	$+0.0038$	$+0.0102$	0.00	3
U-img	U-bSFT-Random	3000	$+0.0034$	$+0.0093$	0.00	3
U-img	U-bSFT-SYNTH	1000	$+0.0471$	$+0.0260$	0.00	3
U-img	U-bSFT-SYNTH	3000	$+0.0544$	$+0.0243$	0.00	3
U-img	U-bSFT-Safety	1000	-0.0005	-0.0006	0.00	3
U-img	U-bSFT-Safety	3000	-0.0001	$+0.0012$	0.00	3
U-img	baseline	0	$+0e+00$	$+0e+00$	0.00	3

Table 8. All-cells systematic table. "Mod." = canary modality (img = saturated image canaries, txt = text canaries, U-img = undertrained image canaries). Δ values are means across 3 seeds, paired-baseline subtraction per canary. Compare against canonical cells C1–C5 in Table 2.

G. Merge-and-unload composition ablation

Stacked probe = base (with merged canary) + bSFT-LoRA loaded; **Merged** probe = base (with merged canary) + bSFT-LoRA merged into the base, then unloaded. Canary memory is intact in both cases; only the bSFT contribution changes from a LoRA delta to a full-rank weight update. Both deltas use the same M_{canary}-only baseline. The ratio merged/stacked indicates how much of the stacked-LoRA drift survives the LoRA-to-full-rank conversion.

The "**Stacked**" column in Table 10 (C3 $+0.0126$, C5 $+0.0055$) comes from the merge-comparison aggregation pipeline; the canonical Δ_{hex} in Table 2 (C3 $+0.0133$, C5 $+0.0076$) comes from the $K=30$ token-level probe pass. The two paths agree in sign at every cell; the magnitude gap (≈ 0.002, max 0.0021 on C5) reflects merge-pipeline aggregation drift, not the bf16 path residual. Table 2 remains the canonical headline estimator; Table 10 uses only stacked-vs-merged *ratios*.

Sanity check: merge does not erase all measured canary-span deltas. Secret-span deltas on C1, C3, C5 survive merging at ratios $+0.98, +0.98, +0.77$; even on C4 itself the secret-span hex delta survives at $+0.51$, so the bSFT delta is not

wholesale erased. Per-token NLL agreement (Table 9) shows mean $|\Delta\text{NLL}_{\text{stacked\_vs\_merged}}| \leq 0.005$ on `canary_hex` for C1–C4, while C4's `preamble` $|\Delta| = 0.02837$ is $\sim 118\times$ its own secret-span $|\Delta| = 0.00024$. The stacked-vs-merged disagreement isolates to C4's preamble tokens, consistent with the Case 2 claim that C4's mean_{20} does not reflect secret-span damage.

Cell	preamble	canary_lit	canary_hex	n tokens
C1	0.00002	0.00000	0.00100	1800
C2	0.00041	0.00002	0.00011	1800
C3	0.00003	0.00000	0.00388	1800
C4	**0.02837**	0.00000	0.00024	1800
C5	0.00001	0.00771	0.04181	1800

Table 9. Mean per-token $|\Delta\text{NLL}|$ between stacked and merged evaluation paths on the same canary captions ($\sim$1800 tokens per cell). C4's preamble $|\Delta|$ is $\sim 118\times$ its own `canary_hex` $|\Delta|$, matching the Mode 2 claim that C4's mean_{20} sits on `preamble` and is composition-sensitive there. C5's larger `canary_hex` number reflects bf16 sensitivity of the undertrained baseline, not Mode 2.

The C4 preamble collapse to $+0.05$ ratio is consistent with several non-exclusive mechanisms (template-token-activation precision sensitivity, BF16 merge rounding for low-magnitude weight updates, LoRA-composition-specific effects) and we make no causal claim. The weak conclusion — a probe whose mass sits on `preamble` under loaded-LoRA evaluation is not robust under merged-LoRA evaluation, even though `canary_hex` deltas on the same and other cells are robust — is sufficient to demote C4's mean_{20} from a memorization claim.

Cell	Span	Stacked	Merged	Merged/Stacked
C1	preamble	$+0.0001$	$+0.0001$	+0.96
	canary_hex	$+0.0029$	$+0.0029$	+0.98
C2	preamble	$+0.0004$	$+3e-05$	+0.07
	canary_hex	$+0.0003$	$+0.0002$	+0.75
C3	preamble	$+6e-05$	$+4e-05$	+0.61
	canary_hex	$+0.0126$	$+0.0124$	+0.98
C4	preamble	$+0.0297$	$+0.0014$	+0.05
	canary_hex	$+0.0004$	$+0.0002$	+0.51
C5	preamble	$+2e-05$	$+1e-05$	+0.81
	canary_hex	$+0.0055$	$+0.0043$	+0.77

Table 10. Merge-and-unload composition ablation across the five canonical cells. M_{canary} is merged into the base before bSFT in both conditions; only the bSFT-LoRA composition differs. **Stacked**: bSFT LoRA loaded as adapter on top of canary-merged base. **Merged**: bSFT LoRA merged into the canary-merged base, then unloaded (no LoRA at probe time). The C4 T-BSFT-SAFETY preamble drift collapses to near zero under merging while C1, C3, C5 secret-span effects survive merging — the C4 signal is evaluation-stack-dependent (loaded vs. merged LoRA) (we do not claim a complete causal mechanism).

H. Direct re-exposure sanity check

To check that the secret-span and behavioural probes *can* move in this testbed (mean20 was the only probe that moved on saturated cells), we trained an additional T-bSFT-GUI variant where the GUI bSFT data was *augmented* with the 20 full-template text-canary captions themselves (full `TEXTCAN-{16hex}-END`, not a 20-token truncation) — explicit re-exposure of the secret content, not a benign-SFT condition. Across 3 seeds: $\Delta\text{NLL}_{\text{mean20}} = +4\times10^{-5}$ (vs. $+9\times10^{-5}$ on plain T-bSFT-GUI; flat both ways), $\Delta\text{NLL}_{\text{hex}} = +0.0006$ (vs. $+0.0133$; $\sim 95\%$ collapse), hit@1 $= 1.00$ (vs. 0.88; recovers to ceiling). The secret-span and behavioural probes move in the recovery direction when secret content is explicitly re-trained on, while mean20 stays flat. This is a probe-sensitivity sanity check, not a Whack-a-Mole-style reactivation control: direct re-exposure is the easiest possible positive contrast, and we use it only to address "can the probes move?".

I. C3 failure inventory and hit@k

For C3 (T-bSFT-GUI step 5000 on text canaries) we report all greedy generations across 20 canaries $\times$ 3 seeds $=$ 60 generations (hit@1) plus $k-1$ stochastic samples at $T = 0.7, p = 0.95$ for hit@4 and hit@16 (i.e. 3 and 15 samples beyond greedy, in the convention defined in §2).

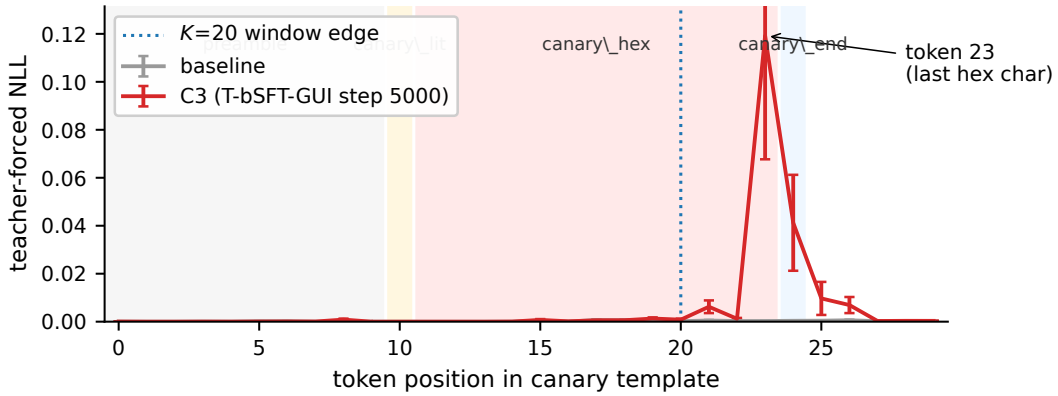

Figure 4. Position-aligned mean per-token NLL across 60 (canary, seed) pairs for C3 vs. baseline. Token 23 (the final `canary_hex` BPE piece, containing the last 1–2 hex characters; 1 for canary 0/14, 2 for canary 12) jumps to 0.119 ($\sim$450$\times$); tokens 0–22 flat. (Per-canary spikes vary; canary 0 alone moves $+0.1228$ on its 13-token hex span, hence canonical $+0.0133$ via per-canary aggregation.)

Seed	**hit**@1	**hit**@4	**hit**@16
seed 0	17/20	18/20	19/20
seed 1	18/20	19/20	19/20
seed 2	18/20	19/20	20/20
total ($N=60$)	53/60	56/60	58/60

Table 11. Per-seed hit@k for C3. The damage is greedy-decode-specific: hit@1 drops below 1.00 on every seed, but the secret is recovered under $k \geq 4$ temperature samples on most or all canaries, indicating the model still has the secret encoded but greedy decoding lands on a near-miss tail.

All 7/60 greedy failures are *wrong hex tail* (last 1–2 characters), e.g. target `TEXTCAN-139bed18a2190bba-END` $\rightarrow$ greedy `TEXTCAN-139bed18a2190b68-END`; this confirms the Case-1 token-level analysis that damage concentrates on trailing hex tokens outside $K=20$.

BPE token disclosure. Qwen2.5's BPE boundaries vary per canary, so the final `canary_hex` piece (token 23) holds the last 1–2 hex characters: 1 for canaries 0/14 (which fail 3/3 seeds), 2 for canary 12 (1/3 seeds). Per-canary string, predicted-token, and rank-of-correct-token manifests are authors' internal artifacts and not released.

Leave-one-canary-out robustness. Headline $\Delta_{\text{hex}} = +0.0133$. Removing canary 0 (its largest single-canary contributor, $+0.123$ on its own span) gives $+0.0075$ ($\sim$57%); the other 19 leave-one-out point estimates lie in $[+0.01332, +0.01400]$ ($<5\%$ of headline). The behavioural break is concentrated in 3/20 canaries with the same Mode 1 token-23 pattern, not a single canary's idiosyncratic tokenisation.

J. Image–bSFT dataset overlap audit

Image canaries draw from ScreenSpot v2 (Wu et al., 2025), a revised release of the original ScreenSpot benchmark of Cheng et al. (2024), and the same source as the GUI bSFT data; a near-duplicate would trivially explain the C1 erosion as in-distribution relearning. We verify no overlap two ways: **(i)** filename-set exclusion of the 20 canary images from bSFT sampling; **(ii)** perceptual-hash check (`phash`, 8×8 DCT, distance ≤ 4) across all canary–bSFT image pairs — no flagged pairs found in 20×5000 comparisons. The text-canary placeholder is a single 224×224 mid-gray bitmap shared by SAFETY and RANDOM by construction; a varied-image text-canary control is future work.

K. Smaller-family decoy audit and Llama-1B trajectory details

This appendix collects the smaller-family-only descriptive evidence that the body summarises but does not print: the decoy table at the candidate C5-equivalent cells (Tab. 12), the migrated Qwen-1.5B C4-equivalent cell, and the Llama-3.2-1B

step-trajectory disagreements. None of this evidence pertains to the original 7B C5 cell, and we read all of it as descriptive smaller-family stress tests (per the caption of Tab. 3).

Table 12. **Decoy probe at the smaller-family candidate C5-equivalent cells (not the original 7B C5 cell).** For each model family, mean Δ relative to the *true* target, averaged across $n = 3$ seeds. If Δ_{mean20} on shuffled-hex / wrong-secret / format-only decoys tracks the Δ on `true`, the in-window NLL drop is format/template smoothing rather than secret-specific learning. The original 7B C5 cell does not have a paired decoy audit and is not represented in this table.

Family	Decoy	$\Delta^{\text{vs. true}}_{\text{mean20}}$	$\Delta^{\text{vs. true}}_{\text{hex}}$	hit@1
Llama-3.2-1B	true	$+0.0000$	$+0.0000$	0.000
Llama-3.2-1B	shuffled_hex	$+4.1514$	$+9.2485$	0.000
Llama-3.2-1B	wrong_secret	$+2.5014$	$+4.8954$	0.000
Llama-3.2-1B	format_only	$+3.4520$	$+10.8553$	0.000
Qwen2.5-1.5B	true	$+0.0000$	$+0.0000$	1.000
Qwen2.5-1.5B	shuffled_hex	$+3.2050$	$+6.4259$	0.000
Qwen2.5-1.5B	wrong_secret	$+3.2336$	$+4.3457$	0.000
Qwen2.5-1.5B	format_only	$+1.6702$	$+2.2722$	0.000

Migrated Qwen-1.5B C4 cell. On Qwen2.5-1.5B the C4-equivalent cell migrates to T-bSFT-Random (5000): $\Delta_{\text{mean20}} = +0.222$, $\Delta_{\text{hex}} = +9 \times 10^{-4}$, hit@1 $= 1.00$; 98% of the mean drift sits on `preamble` with `canary_hex` essentially unchanged — a sharper version of the 7B C4 mechanism.

Llama-3.2-1B step-trajectory disagreements. Two phenomena absent from the 7B testbed appear at the Llama-1B C3/C4-equivalent cells across step $\in \{1k, 3k, 5k\}$: **(i)** on T-bSFT-GUI and T-bSFT-Safety the disagreement *grows* with bSFT magnitude rather than collapsing ($\Delta_{\text{mean20}} = +0.56$ to $+0.81$ at step 5000 co-occurs with greedy hit@1 break 0.37/0.00 while teacher-forced `canary_hex` stays within $1.6\times$ baseline); **(ii)** T-bSFT-GUI shows emergent recovery in greedy hit@16 (seed-mean $0 \to 0.20 \to 0.78$) that is *undetectable* via $\text{mean}_{K=20}$ NLL, which grows monotonically over the same range. Inter-seed variance at step 3000 is high; we report a descriptive trajectory, not a dose-response claim.