WHERE MEMBERSHIP LIVES: TASK ABSTRACTION, EXEMPLAR RETRIEVAL, AND PRIVACY LEAKAGE IN IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning (ICL) creates a different privacy problem than pretrained and fine-tuned language models: adaptation occurs *without* parameter updates, while sensitive demonstrations remain explicitly available in the active context and can be re-accessed during inference. We study this *contextual membership* and ask where membership information resides, how it differs from parameter-updated membership, and how it can be inferred under different model accesses. We introduce controlled behavioral interventions that separate target-specific exemplar information, surface-token sensitivity, and shared task abstraction. A matched ICL–Full-SFT comparison experiment further reveals different privacy footprints: under paraphrased target-associated exposure, ICL yields stronger pooled membership signal than SFT (0.869 versus 0.688) and a differently concentrated token-level signal. Finally, replacing private demonstrations with differentially private synthetic demonstrations reduces the evaluated attack to approximately chance, while increasing prediction loss. Together, these results identify active-context exposure as a distinct source of privacy leakage—different from the persistent parameter traces studied in pretraining and fine-tuning—and motivate privacy analyses and defenses specifically designed for ICL.

1 INTRODUCTION

In-context learning (ICL) and fine-tuning (FT) can adapt the same pretrained language model to the same downstream task, but they store and reuse task information in fundamentally different ways. Fine-tuning changes model parameters, so any example-specific influence must survive compression into a shared parameter update. ICL instead keeps demonstrations explicitly available in the active context, allowing the model to both infer a shared task and re-access individual examples at inference time. This distinction is central to current efforts to understand whether ICL is merely a parameter-free route to similar behavior or a qualitatively different learning mechanism (Xie et al., 2022; Min et al., 2022; Todd et al., 2024; Yin & Steinhardt, 2025; Ghosh et al., 2026). Privacy provides another particularly different perspective on that question: two models can achieve similar task utility while exposing very different evidence about the examples from which they adapted.

Membership inference attacks (MIAs) ask whether a candidate record participated in a model's adaptation process (Shokri et al., 2017; Yeom et al., 2018). For a fine-tuned model, membership refers to a persistent trace in parameters. For ICL, however, the sensitive record can remain verbatim in the prompt and influence the query through contextual computation without any parameter update. Existing attacks establish that LLM membership can be inferred from loss, calibrated likelihood, neighborhood comparisons, tail-token statistics, or text-only behavior (Mattern et al., 2023; Shi et al., 2024; Duan et al., 2024; Wen et al., 2024). Yet a low target loss under ICL is intrinsically ambiguous.

We therefore ask a mechanism-oriented question: *where does membership live in ICL, and how is that footprint different from the privacy trace produced by fine-tuning?* Our central hypothesis is that ICL privacy contains the following three nested information sources. **Task abstraction** captures information shared by demonstrations from the same task and can improve many queries without identifying any particular record. **Exemplar retrieval** captures candidate-specific information that remains available due to the presence of the target exemplar itself, or target-associated evidence in the

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

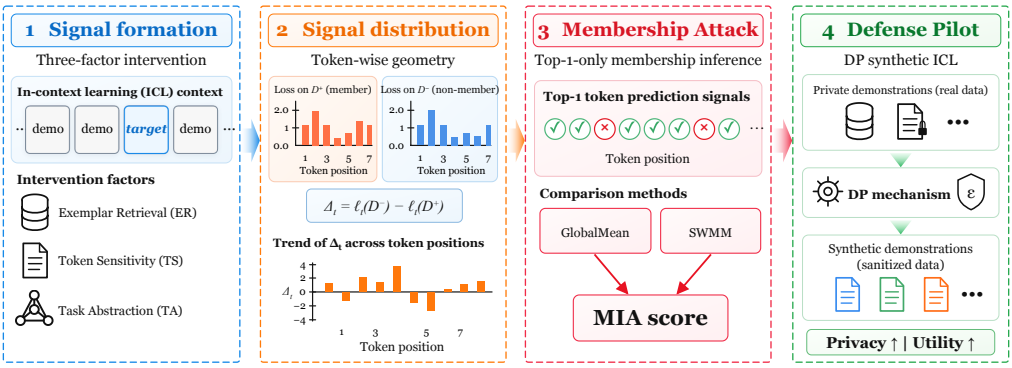


Figure 1: Overview of our main contributions: signal formation, token-wise distribution, top-1-token-only exploitability, and DP synthetic-ICL mitigation. The ICL signal information consists of task abstraction, exemplar retrieval and token sensitivity. The membership signal distribution is the token-wise loss difference between membership and nonmembership. We use the top-1-only prediction method as black-box MIA. We test DP synthetic ICL in our defense experiment.

prompt. **Surface-token sensitivity** captures additional dependence on the exact lexical or symbolic *surface* realization of that exemplar. The latter two sources are the natural carriers of membership, whereas task abstraction is principally a utility-preserving confounder that an MIA should control.

Prior membership-inference work largely asks whether model outputs distinguish members from nonmembers (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022). For ICL, we focus on the following three questions: *Does the signal reflect a shared task relation, target-specific information, or exact-token reuse? Is the signal concentrated in a few output positions or spread across the sequence? Does it remain visible without probabilities or losses?* The answers also determine which observable differences an effective defense must obscure.

We answer these questions through different experiments (Figure 1). A three-factor factorial experiment estimates target-specific exemplar effects, surface-form sensitivity, and task-structure modulation (or task abstraction); a three-dataset analysis characterizes their token-wise geometry. We then evaluate a top-1-token-only attack and perform a differentially private (DP) synthetic-demonstration privacy-preserving experiment. A matched ICL–Full-SFT comparison contrasts active-context exposure with parameter updates. Our main contributions are:

- a controlled behavioral framework that separates target-specific exemplar effects, surface-form sensitivity, and task abstraction;
- cross-domain (formal vs. natural language) evidence that contextual membership has broad but distinct token-level geometry;
- top-1-token-prediction showing that exact contextual membership remains highly separable;
- a preliminary privacy-defense experiment in which the original private ICL demonstrations are replaced by differentially private (DP) synthetic demonstrations, and then evaluates both privacy leakage and model utility.

2 Problem Formulation and Basic Definitions

Let a frozen autoregressive model f_θ receive an ordered demonstration context $D = \{(x_j, y_j)\}_{j=1}^{K}$ and a target $z = (x, y)$, where $y = (y_1, \ldots, y_T)$. Let D^+ be the demonstration set (or context) which contains the target exemplar, let D^- replace it with a matched example from the same task distribution. The target, the remaining demonstrations, their order, and the scoring procedure are held fixed whenever possible. The model uses a private context D hidden from the attacker. Given a known candidate exemplar $z = (x, y)$, the attacker distinguishes $H_1 : z \in D$ from $H_0 : z \notin D$ through repeated teacher-forced queries to the same context. At each position, the API returns only the top-1 predicted token ID, which is compared with the known target token to obtain a correctness bit. It reveals no probabilities, losses, logits, ranks, internal states, gradients, parameters, or view of D. Paired D^+/D^- comparisons are controlled diagnostics. Figure 2 illustrates the paired scoring

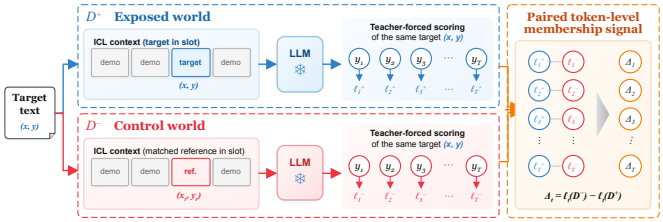


Figure 2: Paired token-level membership signal under exposed D^+ and control D^-.

construction behind the token-level contrast in Section 2. Because this construction uses both actual and counterfactual worlds, it is a behavioral diagnostic rather than a deployable attack. The latter attack observes only one world and assigns only one actual attack score.

At target position t, teacher forcing supplies the true prefix $y_{<t}$ and evaluates the next-token prediction loss as $\ell_t(D) = -\log p_\theta(y_t \mid D, x, y_{<t})$. We define the sequence-level loss as the average over all T target positions, $L(D) = \frac{1}{T}\sum_{t=1}^{T} \ell_t(D)$. To quantify how contextual membership affects prediction at each target position, we define the token-level membership signal as $\Delta_t = \ell_t(D^-) - \ell_t(D^+)$, where D^+ is the exposed world containing the target exemplar and D^- is its matched counterfactual control. A positive signal, $\Delta_t > 0$, means that $\ell_t(D^+) < \ell_t(D^-)$; in other words, exposing the target exemplar makes the ground-truth token at position t easier to predict. We therefore interpret $\Delta_t > 0$ as evidence that token t supports membership, whereas $\Delta_t < 0$ indicates no positive membership advantage at that position. We use Δ_t as the token-level variable in the following behavioral analysis.

3 Related Work

MIA for language models and ICL contextual privacy. Membership inference distinguishes training members from nonmembers using model behavior (Shokri et al., 2017; Yeom et al., 2018). Subsequent work studies likelihood, calibration, neighborhood comparisons, and data extraction for language models (Carlini et al., 2021; 2022; Mattern et al., 2023; Shi et al., 2024; Duan et al., 2024). Their results aimed to establish (pre)training or finetuning-data membership. Demonstration content, labels, and order can shape ICL behavior (Min et al., 2022; Xie et al., 2022). Work on induction heads, task vectors, and function vectors proposes internal accounts of ICL (Olsson et al., 2022; Hendel et al., 2023; Todd et al., 2024; Yin & Steinhardt, 2025). Our prompt interventions instead define behavioral estimands and do not identify these internal mechanisms. Wen et al. (2024) propose four attack strategies tailored to various constrained scenarios and conduct extensive experiments on four popular large language models. FALCON studies text-only membership inference against ICL (Su et al., 2025). WBC aggregates local likelihood evidence for fine-tuned language models (Chen et al., 2026).

Private prompting and complementary datasets. Prior work protects ICL through private aggregation, synthetic few-shot generation, or local DP (Wu et al., 2024; Tang et al., 2024; Zheng et al., 2024). Our DP experiment is a preliminary evaluation of the proposed defense rather than a new formal privacy guarantee. The datasets serve complementary roles across our evaluation pipeline: CLAPNQ (Rosenthal et al., 2025) supports the factorial analysis; Formal L1 (Ghosh et al., 2026) provides both the positive-control setting and the domain for the defense study; all three datasets are used to characterize token-level membership profiles; and the attack evaluation is conducted on Formal L1 and CNN/DailyMail (Hermann et al., 2015). Accordingly, our experiments use different but related settings to address different research questions.

4 Three-Factor Factorial Analysis of Contextual Membership

We operationalize each concept through a controlled prompt change and measure the resulting loss difference:

$$\begin{aligned} \text{Present vs. Replacement} &\longrightarrow \text{Exemplar Retrieval } (\Delta_{\mathrm{ER}}), \\ \text{Original vs. Paraphrased} &\longrightarrow \text{Token Sensitivity } (\Delta_{\mathrm{TS}}), \\ \text{Valid vs. Incoherent} &\longrightarrow \text{Task Abstraction } (\Delta_{\mathrm{TA}}). \end{aligned} \tag{1}$$

The names refer to the concepts being operationalized; the quantities on the right are measurements of model behavior.

Exemplar Retrieval. The Present condition places the target exemplar in a designated prompt slot. Replacement puts a matched non-target exemplar in that slot. The evaluated answer, background, relation condition, and surface condition remain fixed. In Paraphrased conditions, the slot contains the paraphrase of the target or replacement, while scoring still uses the original target answer. This intervention measures the marginal benefit of target-associated evidence relative to the chosen replacement; it does not isolate an internal retrieval operation.

Token Sensitivity. Original preserves the exemplar's question and answer tokens. Paraphrased changes their wording while approximately retaining meaning, in both Present and Replacement conditions. The background relation and evaluated target remain fixed. Comparing the two Exemplar Retrieval contrasts tests dependence on exact token form after subtracting the corresponding replacement baseline.

Task Abstraction. Valid background demonstrations exhibit a consistent input–output relation. Incoherent backgrounds disrupt that relation while retaining the prompt format. Within each surface condition, we compare the Present-minus-Replacement benefit across these backgrounds. This asks whether shared task information changes the marginal value of the target exemplar. A loss benefit shared equally by Present and Replacement cancels in this contrast; a near-zero value therefore does not establish that the model ignores task information.

Let $r \in \{v, c\}$ denote Valid and Incoherent backgrounds, respectively, and $s \in \{o, p\}$ denote Original and Paraphrased exemplars. The index c is only a shorthand for Incoherent. For the same target and background draw, $L_{r,a,s}$ is mean target-answer NLL, where a specifies Present or Replacement. We compute loss difference within these matched units and then aggregate them:

$$\Delta_{\mathrm{ER}}^{(r,s)} = L_{r,\mathrm{replacement},s} - L_{r,\mathrm{present},s}. \tag{2}$$

Positive Δ_{ER} means the target exemplar reduces loss relative to its matched replacement. A larger positive value means a larger reduction in nats/token; a negative value means the replacement yields lower loss. This is a paired exposure effect, not attack AUC.

$$\Delta_{\mathrm{TS}}^{(r)} = \Delta_{\mathrm{ER}}^{(r,o)} - \Delta_{\mathrm{ER}}^{(r,p)}. \tag{3}$$

Positive Δ_{TS} means the Exemplar Retrieval contrast is larger with the original tokens than with paraphrases. A negative value means the contrast is larger after paraphrasing. Its absolute magnitude measures the change in that contrast, in nats/token.

$$\Delta_{\mathrm{TA}}^{(s)} = \Delta_{\mathrm{ER}}^{(v,s)} - \Delta_{\mathrm{ER}}^{(c,s)}. \tag{4}$$

Positive Δ_{TA} means that the target exemplar provides greater predictive benefit under Valid backgrounds, whereas negative Δ_{TA} means that it provides less. The latter can occur when the valid rule already helps predict the target even without its exemplar. Neither sign alone measures task accuracy or identifies an internal Task Abstraction mechanism. We define the three-way interaction as $\Delta_{\mathrm{ER}\times\mathrm{TS}\times\mathrm{TA}} = \Delta_{\mathrm{TA}}^{(o)} - \Delta_{\mathrm{TA}}^{(p)}$, which measures how the Task Abstraction contrast changes across surface conditions. A value of zero means that the two conditional Task Abstraction contrasts coincide; it does not imply that ER, TS, and TA have no effects individually. All of these quantities are conditional contrasts rather than additive shares of total membership leakage.

4.1 Experimental Setup

The natural-language experiment uses 400 CLAPNQ targets, each with four background draws, and the frozen base `Qwen/Qwen2.5-7B` checkpoint. CLAPNQ supplies question–answer examples. For each target and draw, we cross Present/Replacement, Original/Paraphrased, and Valid/Incoherent to obtain $400 \times 4 \times 8 = 12{,}800$ prompt conditions. We score the same original answer in every condition. The model, nonmanipulated prompt components, and scoring path are fixed. Figure 3 shows what changes along each axis. Throughout the factorial results, a value followed by brackets denotes a point estimate followed by its 95% confidence interval (CI). Intervals use 2,000 target-cluster bootstrap draws: a resampled target retains all its background draws and factorial cells.

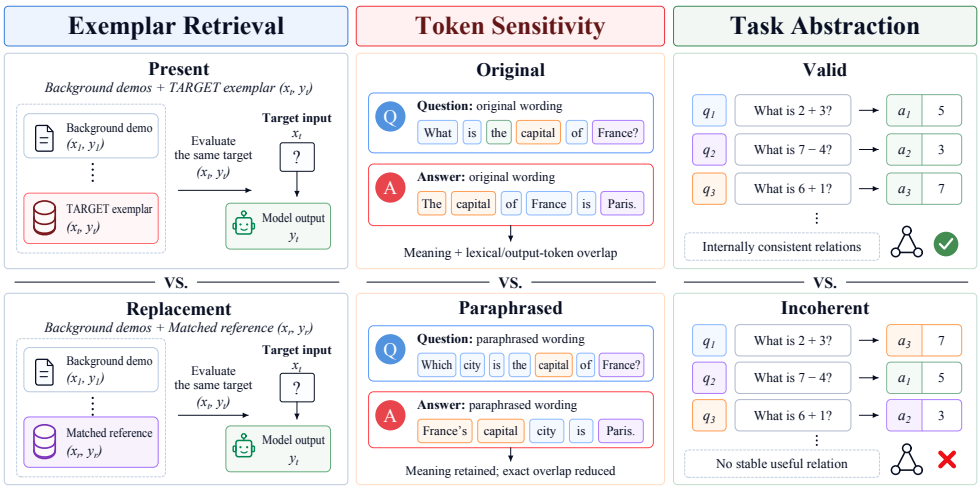


Figure 3: Behavioral three-factor factorial design. The experiment crosses exemplar availability, token overlap, and relation structure while holding the target and background draw fixed.

Primary randomization tests use 20,000 target-level sign flips of paired effects. Thus tokens and repeated prompt conditions are not independent observations. A crossed question/answer probe varies their surface forms separately, adding original-question/paraphrased-answer and paraphrased-question/original-answer cells. A formal positive control adds a Coherent-wrong background with a consistent alternative rule. The supplied plotting documentation identifies the valid formal rule as string reversal. These analyses ask which side of the exemplar contributes to Token Sensitivity and whether Δ_{TA} responds to controlled task structure. Appendix B records construction details.

4.2 Exemplar Retrieval remains positive after paraphrasing

With valid backgrounds and original surface forms, $\Delta_{\mathrm{ER}}^{(v,o)} = 1.9441$ nats/token, with a 95% confidence interval (CI) of $[1.8795, 2.0104]$. After paraphrasing, $\Delta_{\mathrm{ER}}^{(v,p)} = 1.2635$, with a 95% CI of $[1.2020, 1.3284]$. Figure 4A shows that paraphrasing attenuates but does not remove the paired availability effect. The effect persists under a strict joint paraphrase-quality filter (1.1955, with a 95% CI of $[1.0690, 1.3106]$, 93 targets) and with the nearest available same-type replacements (1.8928 original; 1.2442 paraphrased, 200 targets). Thus, the paired loss reduction remains positive after paraphrasing under both the standard and additional replacement checks. This behavioral evidence does not identify a unique internal retrieval mechanism.

4.3 Token Sensitivity and the answer-surface probe

The valid-background token-sensitivity effect is $\Delta_{\mathrm{TS}}^{(v)} = 1.9441 - 1.2635 = 0.6806$ nats/token, with a 95% CI of $[0.6516, 0.7109]$. In the crossed surface probes, the question-surface contrast is -0.0003, with a 95% CI of $[-0.0046, 0.0040]$, whereas the answer-surface contrast is 0.6663, with a 95% CI of $[0.6126, 0.7209]$. The question-surface value is reported as -0.0003 after rounding the signed estimate -0.0003093; it is neither missing nor clipped. The retained aggregates do not support an interval estimate for this residual. The answer-surface estimate of 0.6663 is close in magnitude to the valid-background token-sensitivity effect, $\Delta_{\mathrm{TS}}^{(v)} = 0.6806$, whereas the question-surface confidence interval includes zero (Figure 4B). This pattern is consistent with the effect depending primarily on answer tokens, but it does not establish a causal percentage attributable to copying.

4.4 Task Abstraction differs across natural and formal settings

Figure 5A reports the Task Abstraction contrasts. The original-surface interval includes zero, $\Delta_{\mathrm{TA}}^{(o)} = -0.00005$, with a 95% CI of $[-0.0056, 0.0051]$, whereas the paraphrased-surface interval excludes zero, $\Delta_{\mathrm{TA}}^{(p)} = 0.0111$, with a 95% CI of $[0.0055, 0.0168]$. In Replacement conditions,

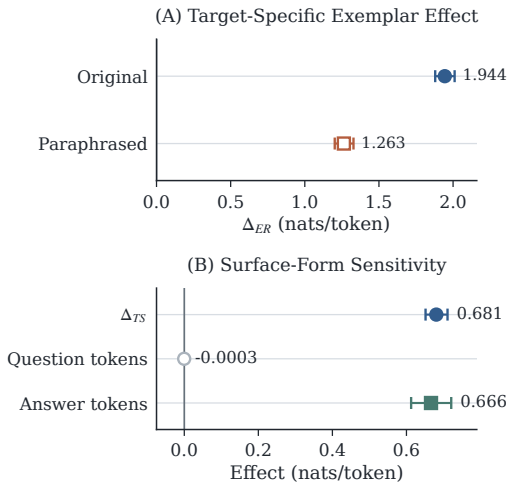


Figure 4: Target-specific effects survive paraphrasing; surface sensitivity is concentrated on answer tokens. Error bars: 95% target-cluster CIs.

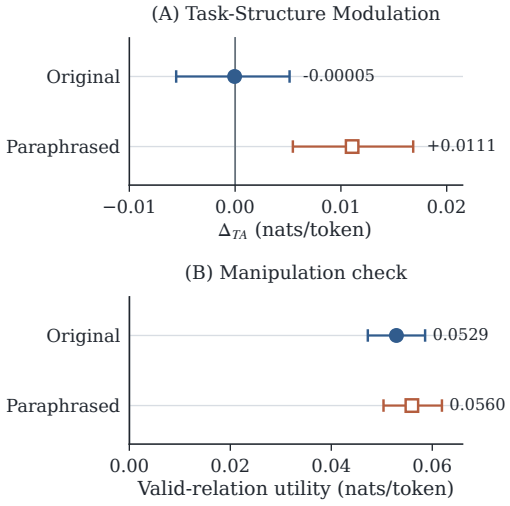


Figure 5: Usable task structure improves utility but weakly modulates Δ_{ER}. Error bars: 95% target-cluster CIs.

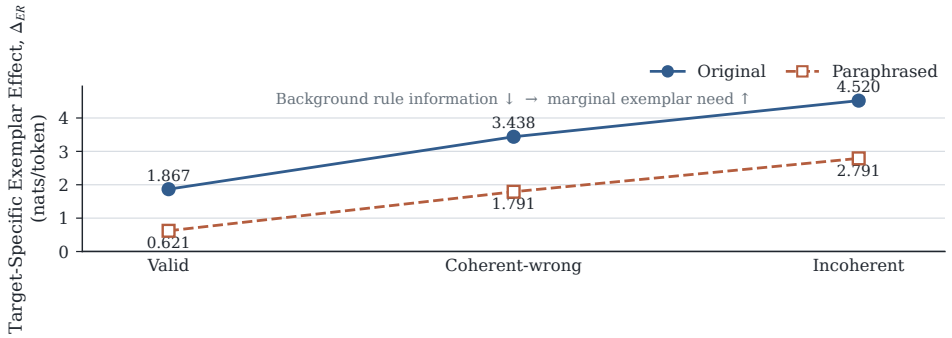


Figure 6: Formal positive control: Δ_{ER} rises as rule information weakens, a pattern consistent with rule–exemplar substitution. Error bars: 95% target-cluster CIs.

valid structure lowers target loss relative to incoherent structure by 0.0529 nats/token, with a 95% CI of $[0.0472,\ 0.0586]$, with original tokens and by 0.0560 nats/token, with a 95% CI of $[0.0503,\ 0.0619]$, after paraphrasing. The manipulation affects prediction, but this benefit appears in both Present and Replacement conditions and largely cancels in Δ_{ER}. The three-way interaction is -0.0111, with a 95% CI of $[-0.0145,\ -0.0079]$, indicating a change in the conditional interaction that is smaller in magnitude than the reported Exemplar Retrieval and Token Sensitivity contrasts. The formal positive control produces Task Abstraction contrasts with greater absolute magnitudes than those in CLAPNQ (Figure 6). For original tokens, Δ_{ER} increases from 1.8671 under Valid structure to 3.4375 under Coherent-wrong structure and 4.5198 under Incoherent structure. For paraphrased tokens, it rises from 0.6210 to 1.7908 and 2.7910. Relative to Incoherent backgrounds, $\Delta_{\mathrm{TA}}^{(o)} = -2.6527$ and $\Delta_{\mathrm{TA}}^{(p)} = -2.1700$. This negative Δ_{TA} is consistent with *rule–exemplar substitution*: when the background exposes a deterministic rule, the missing target remains predictable, so the target exemplar adds less marginal information. The result shows that the estimand responds to known structure, not that formal and natural tasks share an internal mechanism.

In the Valid/Original CLAPNQ condition, $\Delta_{\mathrm{ER}} = 1.9441$, compared with $\Delta_{\mathrm{TS}} = 0.6806$ and $\Delta_{\mathrm{TA}} = -0.00005$ nats/token. These measurements answer different conditional questions; their magnitudes do not establish a decomposition of membership leakage. These conditional contrasts are not independent causal shares. Table 2 reports all primary contrasts and intervals.

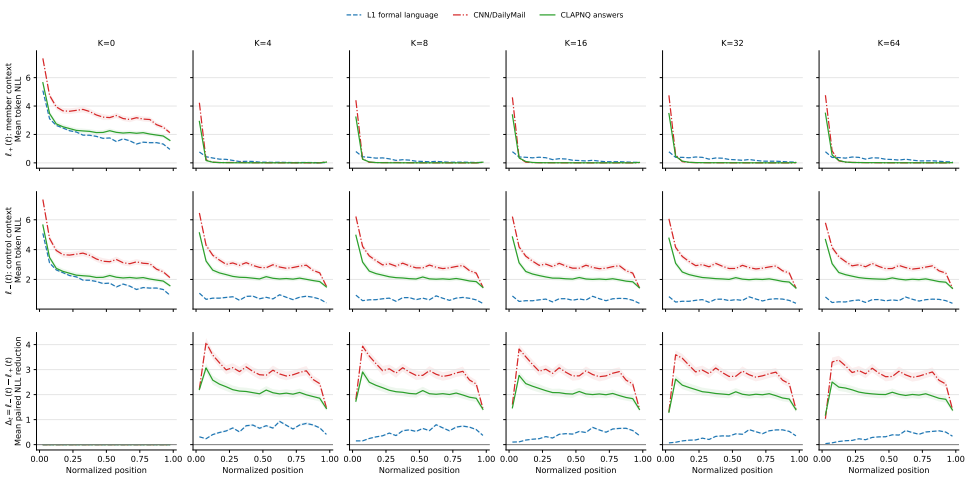


Figure 7: Token-wise contextual-membership geometry for Formal L1, CNN/DailyMail, and CLAPNQ. Columns vary the number of demonstrations; rows show exposed loss, matched-control loss, and paired loss reduction Δ_t.

5 Token-Level Distribution of Membership Signals

We next fix exact exemplar presence across Formal L1, CNN/DailyMail, and CLAPNQ. Formal L1 supplies grammar-generated strings, CNN/DailyMail supplies summary highlights, and CLAPNQ supplies long-form answers. This experiment scores those sequences; it does not evaluate summarization or question-answering task accuracy. For every $K \in \{4, 8, 16, 32, 64\}$, D^+ contains the candidate, whereas D^- places a distinct exact-token-length-matched reference in the same slot. Each dataset uses 500 candidate/reference pairs and three context seeds. The exact candidate occupies the middle slot ($K//2$); the control is a distinct sequence with the same token length. Backgrounds, order, and scoring remain fixed within the pair. We use the same fixed version of the base Qwen2.5-7B model throughout the experiments, without updating its parameters, and fix the exact model revision to ensure reproducibility (Yang et al., 2024). Figure 7 shows the absolute losses and the paired token-level membership signal Δ_t. For natural text, exposed losses drop sharply after an identifying prefix, whereas matched-control losses remain high. CNN/DailyMail and CLAPNQ have positive Δ_t on 98.4–99.6% of tokens. Formal controls benefit from the shared grammar, so their exposure gap is smaller and decreases more with context length. From $K = 4$ to 64, mean Δ falls from 0.6241 to 0.3261 for Formal L1 (-47.7%), from 2.8672 to 2.6970 for CNN/DailyMail (-5.9%), and from 2.1333 to 1.9708 for CLAPNQ (-7.6%). We do not find a single suffix of the target sequence that contains most of the membership information. Instead, the difference in prediction likelihood between member and non-member contexts remains visible across many token positions. We therefore test whether this membership signal can still be detected when the model output is reduced to a simple binary indicator of whether the top-1 predicted token is correct.

6 Top-1-Token-Only Contextual-Membership Inference and Privacy Pilot

The attack is evaluated on Formal L1 and CNN/DailyMail with $K \in \{4, 8, 16, 32, 64\}$. At token position t, the model receives the hidden active context and the ground-truth prefix $y_{<t}$, while the API returns only its top-1 predicted token $\hat{y}_t$. The attacker compares this prediction with the known target token y_t and records the correctness bit $b_t = \mathbb{I}[\hat{y}_t = y_t]$. Each hidden context is scored independently: the attacker never observes both the exposed context D^+ and the control context D^- and therefore cannot use their difference. The access restrictions defined in Section 2.1 therefore remain unchanged. Although the attacker observes only top-1 predictions, obtaining the full sequence of correctness bits requires repeated teacher-forced queries across the target sequence; this is therefore neither a single-query attack nor necessarily a low-cost one.

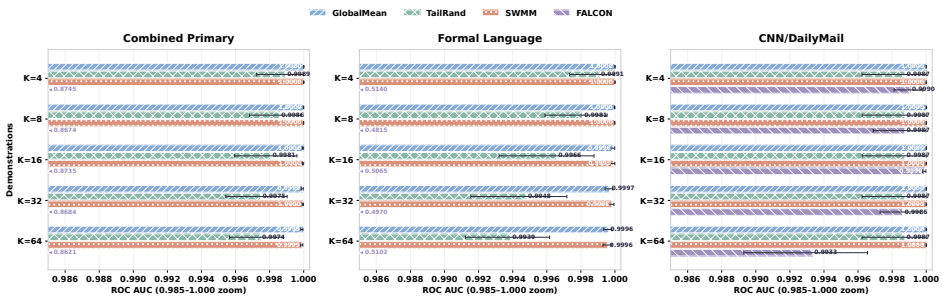


Figure 8: Held-out ROC-AUC versus K for the top-1-token-only methods (95% candidate-cluster bootstrap CIs). FALCON uses a different observation channel, so this is not a like-for-like comparison; off-range values are annotated at the boundary.

We convert the token-level correctness bits into a single membership score using three aggregation methods. GlobalMean-Match averages correctness over the entire target sequence: $S_{\mathrm{GM}} = \frac{1}{T}\sum_{t=1}^{T} b_t$. TailRand-Match randomly selects $m = 16$ positions from the latter half of the sequence ($q = 0.5$) and averages their correctness bits. Our Sliding-Window Majority Match (SWMM) instead summarizes correctness within neighboring token positions: $S_{\mathrm{SWMM}} = \frac{1}{T-w+1}\sum_{u=1}^{T-w+1} \mathbb{I}\left[2\sum_{t=u}^{u+w-1} b_t > w\right]$. SWMM is motivated by the locality of loss-based WBC (Chen et al., 2026), but uses only top-1 correctness bits rather than likelihoods or differences between paired contexts. For each dataset, we use 100 development candidates (examples used to choose the attack hyperparameter before the final evaluation.) to select w from the prespecified range 2–32, and then fix $w = 2$ for 400 separate evaluation candidates. When $w = 2$, a window is counted as correct only when both adjacent tokens are predicted correctly. Thus, SWMM reduces to the fraction of adjacent token pairs that are both correct. Appendix D reports the complete comparison used to choose the best window width.

6.1 Attack Results

On the pooled evaluation set, GlobalMean achieves an AUC of 0.999967 and a strict TPR@1%FPR of 0.999250; SWMM achieves 0.999973 and 0.999583, respectively; and TailRand achieves 0.998110 and 0.943000. FALCON (Su et al., 2025), which uses text perturbations and different attack features, achieves an AUC of 0.869157 and a TPR@1%FPR of 0.503083. To estimate uncertainty, we resample complete candidates 2,000 times while keeping all measurements associated with the same candidate—including all K values, context seeds, and member/control conditions—together. For evaluation at low false-positive rates, candidates with identical attack scores are always kept together rather than split across a decision threshold. GlobalMean and SWMM perform nearly identically, with both close to the performance ceiling. Thus, exact contextual membership remains almost perfectly distinguishable even when the attacker observes only top-1 token predictions, without access to probabilities or losses. However, these results do not show that local aggregation is inherently better than global averaging. TailRand shows that 16 positions sampled from the latter half of the target sequence still contain strong membership information, although its limited number of possible scores reduces its effectiveness at very low false-positive rates. FALCON achieves an AUC of 0.997940 on CNN/DailyMail but only 0.501818 on Formal L1. One possible explanation is that its text perturbations and learned similarity features are less suitable for symbolic sequences, although our experiment does not directly support this explanation. Because FALCON and our top-1 methods rely on different types of model access, these results should not be interpreted as a direct comparison under identical attack conditions (Appendix D).

6.2 Privacy Defense

As a preliminary defense experiment, we replace each private Formal L1 demonstration with a 48-token synthetic string generated under differential privacy (DP). At each generation step, we average the private next-token distributions, add calibrated Gaussian noise, and select the next token.

Table 1: Formal L1 DP-synthetic-ICL development pilot using single-context GlobalMean. AUC intervals use 2,000 candidate-cluster bootstrap draws. Loss is held-out token NLL in nats/token; lower is better.

K	Condition	MIA AUC [95% CI]	Grammar acc.	Loss ↓
8	Raw	1.0000 [1.0000,1.0000]	100.0%	0.703
8	DP $\epsilon = 1$	0.5002 [0.4997,0.5007]	93.4%	2.050
8	DP $\epsilon = 5$	0.5010 [0.4998,0.5025]	93.8%	2.037
8	DP $\epsilon = 10$	0.4997 [0.4979,0.5013]	94.8%	2.027
16	Raw	1.0000 [1.0000,1.0000]	100.0%	0.658
16	DP $\epsilon = 1$	0.5000 [0.4996,0.5003]	92.4%	2.078
16	DP $\epsilon = 5$	0.5003 [0.4981,0.5032]	93.7%	2.059
16	DP $\epsilon = 10$	0.4996 [0.4971,0.5017]	94.4%	2.027

We use one-record replacement as the neighboring-dataset definition, set $\delta = 10^{-5}$, and consider privacy levels $\epsilon \in \{1, 5, 10\}$. The privacy calculation accounts jointly for all $K \times 48$ generated tokens that form the synthetic prompt. We evaluate this defense on the 100 candidates previously reserved for choosing or checking method settings, using $K = 8$ and $K = 16$. The reported ϵ values are the privacy levels specified by the DP accounting procedure. The empirical attack AUC is reported separately and it does not verify that the DP implementation itself is correct.

Across all tested DP settings, GlobalMean AUC drops from 1.0 to 0.4996–0.5010. Grammar discrimination remains 92.4–94.8%, while held-out token NLL increases from 0.658–0.703 to 2.027–2.078. Thus, for the attacks evaluated here, the DP synthetic demonstrations make member and non-member examples nearly indistinguishable, while preserving good performance on the grammar test but substantially reducing sequence-prediction quality. The small AUC differences across ϵ are all close to chance and should not be interpreted as showing a consistent privacy ordering. Appendix E reports the implementation details and the corresponding width-8 likelihood-window results.

7 Conclusion

In CLAPNQ, including the target exemplar in the ICL context produces the largest measured effect, and preserving its exact answer tokens further strengthens this effect. The token-level analysis shows that the target exemplar changes top-1 predictions across many token positions, which is consistent with the strong membership separability observed when the exact target is present. In the separate Formal L1 defense experiment, DP synthetic demonstrations reduce membership separability but increase prediction loss. Because these stages use complementary datasets to answer different questions, they provide related evidence rather than a single end-to-end causal experiment. Across 180 matched Mistral-7B comparisons per method, ICL achieves a higher AUC (0.869 versus 0.688) and has a larger fraction of token positions with a positive membership signal than Full-SFT. However, a larger fraction of the total signal is concentrated in the strongest 10% of token positions, and the signal is spread over fewer positions overall. Thus, in this setting, the ICL signal is positive at more positions but is stronger in a smaller subset of positions, so it is not necessarily more broadly distributed than the SFT signal. Appendix F provides experimental setup and confidence intervals.

Our membership label indicates whether a candidate exemplar is present in the hidden ICL context. When the exact target exemplar is present, attack performance is close to its maximum, and the attack requires a known target and repeated queries. The main experiments use only one fixed version of the Qwen2.5-7B model; results on formal-language strings may not fully generalize to natural-language tasks; and paraphrases preserve meaning only approximately. CLAPNQ is not included in the attack evaluation, while FALCON uses a different type of model access. The Formal L1 defense experiment is limited in scale and its DP privacy calculation still requires independent verification. Finally, the ICL–SFT comparison uses five random runs, one model family, one domain, and fine-tunes the model for only one full pass over the training data.

AI USE STATEMENT

Generative AI tools were used for code assistance, experiment documentation, figure preparation, and language editing. The authors verified generated code, experimental outputs, statistical analyses, and numerical claims and remain fully responsible for the scientific content.

ETHICS STATEMENT

Membership inference can support privacy auditing, but it can also reveal whether a model context contains sensitive evidence. We report aggregate results on research datasets, distinguish diagnostic from operational access, and do not attempt to identify individuals. Released materials should omit private prompts and preserve the access restrictions attached to audit manifests.

REPRODUCIBILITY STATEMENT

The appendix records constructions, estimands, pairing, uncertainty units, pinned model revisions, attack selectors, defense accounting settings, and audit boundaries. Missing uncertainty is labeled as unavailable rather than reconstructed from plots or point estimates.

REFERENCES

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy*, pp. 1897–1914. IEEE, 2022. doi: 10.1109/SP46214.2022.9833649.

Yuetian Chen, Yuntao Du, Kaiyuan Zhang, Ashish Kundu, Charles Fleming, Bruno Ribeiro, and Ninghui Li. Window-based membership inference attacks against fine-tuned large language models. *arXiv preprint arXiv:2601.02751*, 2026. URL https://arxiv.org/abs/2601.02751.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=av0D19pSkU.

Bishwamittra Ghosh, Soumi Das, Till Speicher, Qinyuan Wu, Mohammad Aflah Khan, Deepak Garg, Krishna P. Gummadi, and Evimaria Terzi. Fine-tuning vs. in-context learning in large language models: A formal language learning perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 41651–41697. Association for Computational Linguistics, 2026.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1693–1701, 2015.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.719. URL https://aclanthology.org/2023.findings-acl.719/.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.759/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Sara Rosenthal, Avirup Sil, Radu Florian, and Salim Roukos. CLAPnq: Cohesive long-form answers from passages in natural questions for RAG systems. *Transactions of the Association for Computational Linguistics*, 13:53–72, 2025. doi: 10.1162/tacl_a_00729. URL https://aclanthology.org/2025.tacl-1.3/.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy*, pp. 3–18. IEEE, 2017. doi: 10.1109/SP.2017.41.

Haitao Su, Yue Qin, Zhenhua Li, Xin Miao, and Yuan Zhou. FALCON: A universal text-only membership inference attack framework against in-context learning. *IEEE Transactions on Information Forensics and Security*, 20:12964–12979, 2025. doi: 10.1109/TIFS.2025.3640876.

Xinyu Tang, Richard Shin, Huseyin A. Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oZtt0pRnOl.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against in-context learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 3481–3495, 2024.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sDuCndtmFg.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium*, pp. 268–282. IEEE, 2018. doi: 10.1109/CSF.2018.00027.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72428–72461. PMLR, 2025. URL https://proceedings.mlr.press/v267/yin25e.html.

Chunyan Zheng, Keke Sun, Wenhao Zhao, Haibo Zhou, Lixing Jiang, Shaoyang Song, and Chunlai Zhou. Locally differentially private in-context learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 10686–10697. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.935/.

A Definitions and Estimands

Three distinct membership variables. For an ordered ICL context D, strict exemplar membership is $m_{\mathrm{ex}}(e, D) = \mathbf{1}[e \in D]$. Contextual-evidence membership is $m_{\mathrm{ev}}(c, D) = \mathbf{1}[c \in D]$, where evidence c may support the evaluated target without being identical to it. Training membership is $m_{\mathrm{train}}(z, \mathcal{A}) = \mathbf{1}[z \in \mathcal{A}]$ for a parameter-updating dataset $\mathcal{A}$. The first two variables concern an active prompt; the third concerns the data used to update model parameters. Original/Present and the exact-exposure attack use strict exemplar membership. Paraphrased/Present uses target-associated evidence while still scoring the original target. In Full-SFT, membership concerns the parameter-updating dataset; in the defense pilot, it concerns the private dataset supplied to synthesis. We do not treat these variables as interchangeable.

Paired diagnostics versus operational attacks. The paired quantities Δ_t, Δ_{ER}, Δ_{TS}, and Δ_{TA} compare matched D^+ and D^- worlds for the same target. They estimate the behavioral effects of controlled prompt changes. A single-context attack instead maps one observed world to a score and predicts an unknown membership label. The top-1 attack never receives a difference between paired worlds because only one of them is observable.

Factors and behavioral estimands. We use the same concepts and contrasts as in the main text: Present versus Replacement probes Exemplar Retrieval (Δ_{ER}); Original versus Paraphrased probes Token Sensitivity (Δ_{TS}); and Valid versus Incoherent probes Task Abstraction (Δ_{TA}). Thus, Δ_{ER} is always a Present-versus-Replacement contrast; Δ_{TS} is always the Original-versus-Paraphrased change in Δ_{ER}; and Δ_{TA} is always the Valid-versus-Incoherent change in Δ_{ER}. These are mechanism-motivated behavioral estimands, not identified internal operations.

Attack metrics. For member scores $\{a_i^+\}$ and control scores $\{a_j^-\}$,

$$\mathrm{AUC} = \frac{1}{n_+ n_-} \sum_{i,j} \left(\mathbf{1}[a_i^+ > a_j^-] + \tfrac{1}{2}\mathbf{1}[a_i^+ = a_j^-] \right). \tag{5}$$

Here n_+ and n_- count member and control scores, respectively, and larger scores favor membership. AUC averages all member–control comparisons, counting a tie as one half. Strict TPR@αFPR is the largest fraction of members selected by a threshold whose selected-control fraction is at most α; it includes or excludes each complete threshold group and never splits tied scores. With n_- controls, the nonzero empirical FPR grid has resolution $1/n_-$; we interpret low-FPR results at this resolution.

Signal geometry. For the paired token-level membership signal Δ_t, defined as the token-loss reduction under the exposed context relative to the matched-control context, directional coverage is $R_{\mathrm{pos}} = T^{-1} \sum_t \mathbf{1}[\Delta_t > 0]$. With $p_t = |\Delta_t| / \sum_j |\Delta_j|$, effective support is $S_{\mathrm{eff}} = N_{\mathrm{eff}}/T$, where $N_{\mathrm{eff}} = \exp(-\sum_t p_t \log p_t)$. The top-$q$ absolute share is

$$C_q = \frac{\sum_{t \in \mathrm{Top}_q(|\boldsymbol{\Delta}|)} |\Delta_t|}{\sum_t |\Delta_t|}, \tag{6}$$

and normalized entropy is $H_{\text{norm}} = -\sum_t p_t \log p_t / \log T$. Here Top_q selects the fraction q of positions with the largest absolute signals. Larger R_{pos} means exposure helps at more positions; larger C_q means more magnitude is concentrated in those selected positions. Larger entropy or effective support means the absolute signal is distributed more evenly. These quantities describe different properties and are reported separately.

B Three-Factor Factorial Experiment: Construction and Full Results

B.1 Configuration and construction

We use a fixed version of the base Qwen/Qwen2.5-7B model at revision `d149729398750b98c0af14eb82c78cfe92750796`, without updating its parameters. For the natural-language experiment, we use CLAPNQ with 400 target examples and four independently sampled sets of background demonstrations for each target. For every target and background set, we construct both Valid and Incoherent contexts with the same number of demonstrations and the same prompt layout. The designated prompt slot contains either the target exemplar or a matched same-task replacement, using either its Original or Paraphrased form. This gives $400 \times 4 \times 8 = 12{,}800$ evaluated prompt conditions.

For the stricter paraphrase-quality subset, both the Present and Replacement examples must have semantic similarity of at least 0.80, differ in token length by at most two tokens, preserve all numeric information, and have lexical overlap of at most 0.90. Ninety-three targets satisfy all criteria. For a more difficult replacement comparison, we choose the most similar same-task non-target example available for 200 targets. We also separately vary the question and answer wording by evaluating the original–original, paraphrased–paraphrased, original-question/paraphrased-answer, and paraphrased-question/original-answer combinations. In the formal-language experiment, we add a condition in which the background demonstrations follow a consistent but incorrect input–output rule. This extended formal experiment contains 19,200 conditions, and the complete natural-language and formal records together contain 36,320 entries.

B.2 Paired estimation

We form all conditional effects within each target i and draw d before aggregation. To obtain confidence intervals, we resample complete targets 2,000 times while retaining all associated draws and factorial cells. Primary randomization tests use 20,000 target-level sign flips, giving a minimum reportable value of $1/20001 \approx 0.00005$. We never treat tokens or factorial rows as independent observations.

Table 2: Primary CLAPNQ contrasts in nats/token. Rows list the conditional estimands; columns give estimates, 95% target-cluster bootstrap CIs, and sign-flip randomization p-values. Here v/c means Valid/Incoherent and o/p means Original/Paraphrased. Signs follow Equations 2–4; a larger contrast is not automatically better task performance.

Contrast	Estimate	95% target-cluster CI	randomization p
$\Delta_{\text{ER}}^{(v,o)}$	1.9441	[1.8795, 2.0104]	0.00005
$\Delta_{\text{ER}}^{(v,p)}$	1.2635	[1.2020, 1.3284]	0.00005
$\Delta_{\text{TS}}^{(v)}$	0.6806	[0.6516, 0.7109]	0.00005
$\Delta_{\text{ER}}^{(c,o)}$	1.9442	[1.8753, 2.0122]	0.00005
$\Delta_{\text{ER}}^{(c,p)}$	1.2524	[1.1940, 1.3162]	0.00005
$\Delta_{\text{TA}}^{(o)}$	-0.00005	$[-0.0056, 0.0051]$	0.9849
$\Delta_{\text{TA}}^{(p)}$	0.0111	[0.0055, 0.0168]	0.00015
$\Delta_{\text{ER}\times\text{TS}\times\text{TA}}$	-0.0111	$[-0.0145, -0.0079]$	0.00005

For the first 20% of answer positions, $\Delta_{\text{ER}}^{(v,o)} = 2.1200$ (95% CI: [2.0038, 2.2341]) and $\Delta_{\text{ER}}^{(v,p)} = 1.4236$ (95% CI: [1.3168, 1.5287]) nats/token, using the target-cluster bootstrap described above.

Table 3: CLAPNQ manipulation, surface, and robustness checks in nats/token, with 95% target-cluster bootstrap CIs. The first two rows give Incoherent-minus-Valid loss in Replacement contexts; positive values mean Valid backgrounds improve prediction. Surface rows are separate within-Present probes, not additive shares of Δ_{TS}. Remaining rows test Exemplar Retrieval under stricter matching; n counts targets.

Check	Estimate	95% CI
valid over incoherent utility, replacement/original	0.0529	[0.0472, 0.0586]
valid over incoherent utility, replacement/paraphrased	0.0560	[0.0503, 0.0619]
question surface, valid/present	-0.0003	$[-0.0046, 0.0040]$
answer surface, valid/present	0.6663	[0.6126, 0.7209]
question surface, incoherent/present	0.0123	[0.0081, 0.0166]
answer surface, incoherent/present	0.6737	[0.6204, 0.7276]
strict-quality $\Delta_{\mathrm{ER}}^{(v,p)}$ ($n = 93$)	1.1955	[1.0690, 1.3106]
present–hard replacement, original ($n = 200$)	1.8928	[1.7962, 1.9940]
present–hard replacement, paraphrased ($n = 200$)	1.2442	[1.1560, 1.3380]

Teacher forcing progressively reveals the answer prefix, which may reduce the marginal contribution of context at later positions. This interpretation is positional, not a causal claim about decoding.

For the formal positive control, the Valid-minus-Incoherent Task Abstraction contrast is -2.6527 for original surface forms and -2.1700 for paraphrased surface forms. The corresponding Valid-minus-Coherent-wrong contrast is -1.5705 and -1.1698, respectively. Confidence intervals for these differences are unavailable because the retained aggregate results do not support paired target-level resampling. Figure 6 instead shows intervals for the individual Δ_{ER} estimates, as recorded in the supplied plotting CSV. Those marginal intervals do not determine the uncertainty of their differences.

C Three-Dataset Token-Wise Experiment

C.1 Common protocol

For each nonzero $K \in \{4, 8, 16, 32, 64\}$, each dataset uses 500 candidate/reference pairs and context seeds 0, 1, and 2. D^+ places the exact candidate at slot $K/2$, whereas D^- substitutes a distinct exact-token-length-matched reference. Background identities, order, separator, target, tokenizer, and scoring path remain unchanged. One full-sequence forward pass produces the teacher-forced NLL of every candidate token.

Formal L1 uses the formaLLM grammar `pcfg_cfg3b_disjoint_terminals`; each candidate/reference pair is also matched in grammar probability, derivation depth, and the number of grammar rules used. CNN/DailyMail uses version 3.0.0 highlights. Its overall token-length distribution is matched to that of Formal L1, and each candidate/reference pair is closely matched in sentence, character, and word counts. CLAPNQ uses the cohesive long-form answer from each answerable example and ensures that the examples used as background demonstrations, candidates, matched references, and development examples do not overlap.

Token curves use 20 bins over normalized token positions. For each candidate, we first average the results across context seeds and then aggregate across candidates; uncertainty bands are obtained by resampling complete candidates 2,000 times. Because $K = 0$ contains no candidate exemplar in the context, $\Delta_t = 0$ by construction. We therefore use $K = 0$ only as a reference for absolute loss and exclude it when analyzing how the membership signal changes with the number of demonstrations.

The WBC column in Table 4 reports a local loss-based measure computed from the paired D^+ and D^- contexts; it is not the single-context SWMM attack. The AUC column also uses a loss-based score obtained from one context. These results summarize how strongly exact target presence can be detected when loss information is available, but they do not replace the more restrictive attack evaluation in Section 6 that uses only top-1 predicted tokens.

Table 4: Exact-source-presence results by dataset and K. ℓ_+/ℓ_- are exposed/control NLL and Global is their paired difference, in nats/token; positive Global favors exposure. WBC is a paired local loss diagnostic. R_{pos} is the positive-token fraction and S_{eff} the effective-support fraction (Appendix A). AUC measures single-context loss-based separability; higher means more leakage. L1 and CNN/DM abbreviate Formal L1 and CNN/DailyMail.

Dataset	K	ℓ_+	ℓ_-	Global	WBC	R_{pos}	S_{eff}	AUC
L1	4	.1455	.7696	.6241	1.2026	.9133	.4199	1.00000
	8	.1959	.6940	.4981	1.0015	.8341	.3968	1.00000
	16	.2345	.6486	.4141	.8668	.7573	.3816	1.00000
	32	.2632	.6198	.3566	.7632	.7246	.3738	1.00000
	64	.2727	.5988	.3261	.7117	.7213	.3637	.99997
CNN/DM	4	.2543	3.1215	2.8672	4.9059	.9849	.6199	1.00000
	8	.2641	3.0769	2.8128	4.8225	.9852	.6132	1.00000
	16	.2753	3.0583	2.7830	4.7835	.9913	.6089	1.00000
	32	.2973	3.0371	2.7398	4.7264	.9947	.6044	1.00000
	64	.3192	3.0162	2.6970	4.6844	.9956	.6003	1.00000
CLAPNQ	4	.1907	2.3240	2.1333	4.1117	.9843	.5327	.99946
	8	.2127	2.2867	2.0739	4.0071	.9874	.5301	.99922
	16	.2282	2.2626	2.0345	3.9457	.9898	.5272	.99898
	32	.2426	2.2400	1.9974	3.8913	.9906	.5265	.99859
	64	.2529	2.2236	1.9708	3.8526	.9912	.5250	.99822

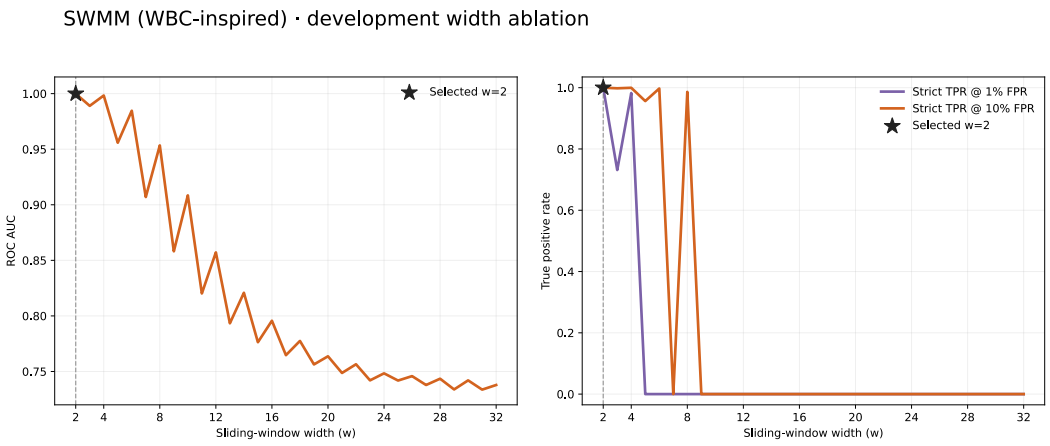


Figure 9: Development-only SWMM width ablation. The horizontal axis is window width w. Left: ROC-AUC; right: strict TPR at 1% (purple) and 10% (orange) FPR. Higher values mean greater attack success. Stars mark the frozen $w = 2$ choice. These development curves guide selection and are not held-out results; no uncertainty bands are shown.

D Top-1-Token-Only Contextual-Membership Attack

D.1 Development and Evaluation Separation

Each dataset has 100 development candidates and 400 disjoint evaluation candidates. On the development set, we select the SWMM width from $w \in \{2, \ldots, 32\}$ and fix TailRand at $q = 0.5, m = 16$. Width $w = 2$ performs best within this prespecified grid and is frozen before evaluation; we do not claim that it is theoretically optimal. The candidate-cluster bootstrap uses 2,000 draws and keeps all K values, context seeds, and worlds for each candidate together. Formal L1 and CNN/DailyMail use the exact-presence construction in Appendix C; the model remains frozen.

D.2 Held-out Aggregate Results

On the combined held-out evaluation set, TPR at 1% FPR is 0.999250 for GlobalMean, 0.943000 for TailRand, 0.999583 for SWMM, and 0.503083 for FALCON. TailRand produces only about 16 distinct possible scores across the evaluation examples. This small number of score values has

Table 5: Held-out attack AUCs aligned on candidate/context keys. Rows are methods; columns pool both datasets or report each separately across K and seeds. Higher AUC means greater membership separability, with 0.5 denoting chance ranking. FALCON uses a different observation channel (Table 6).

Attack	Combined AUC	Formal L1 AUC	CNN/DM AUC
GlobalMean-Match	0.999967	0.999870	1.000000
TailRand-Match	0.998110	0.996540	0.998700
SWMM ($w = 2$)	0.999973	0.999890	1.000000
FALCON	0.869157	0.501818	0.997940

little effect on overall AUC, but reduces the attack's ability to identify members when the allowed false-positive rate is only 1%. Figure 8 reports AUC for each K with uncertainty intervals obtained by resampling complete candidates; annotations show FALCON values that fall outside the displayed axis range.

Table 6: Observation channels in the attack comparison. Rows identify methods and columns identify information channels; a checkmark denotes use and a dash non-use in the reported setup. FALCON is an ICL baseline with text perturbations and learned features, so the performance comparison does not hold access fixed.

Method	Top-1 IDs	Prob./loss	Text perturbation	Learned features
GlobalMean / TailRand / SWMM	✓	–	–	–
FALCON	–	–	✓	✓

The experiment shows that exact target presence changes the model's actual top-1 next-token predictions, rather than only changing their probabilities. However, because GlobalMean and SWMM both perform nearly perfectly, these results do not show that SWMM is clearly better than GlobalMean. They also do not imply that top-1-token-only access is inexpensive, or that FALCON and SWMM assume the same attacker capabilities.

E DP Synthetic-ICL Defense Pilot

E.1 Protocol and accounting boundary

The defense experiment uses the same 100 Formal L1 candidates previously used for method development, with $K \in \{8, 16\}$ and context seeds 0, 1, and 2. For each setting, we evaluate both the member and matched-control context for every candidate and seed, giving $100 \times 3 \times 2 = 600$ observations. Across the Raw baseline and three DP privacy settings, this gives 4,800 complete observations with no duplicated or missing entries. We do not report $K = 32$ because that run stopped after only 61 of the planned 2,400 observations.

Fixed-Gaussian-KV-v2 generates 48 synthetic tokens for each private demonstration using the public digit vocabulary 1–9. At each generation step, it averages the private next-token distributions, adds calibrated Gaussian noise, and selects the next token. Two private datasets are treated as neighbors if they differ by one replaced demonstration. We use $q = 1$, $\delta = 10^{-5}$, and a PRV accountant to compute an upper bound on ϵ. The privacy calculation jointly accounts for 384 token releases at $K = 8$ and 768 at $K = 16$. We use Transformers 4.45.2 with the legacy KV cache; Gaussian noise is assigned to token IDs in ascending order, and the exact DP generation code is fixed at commit `7c3c5dc`. The experiment code does not apply an additional clipping step after private aggregation. Thus, the privacy guarantee relies on the sensitivity bound implemented in that aggregation routine and should be independently verified.

The defense protocol specifies GlobalMean as negative mean candidate-token NLL, $-L(D)$: larger scores indicate that the candidate is easier to predict. The supplementary WBC-window score is the maximum negative mean NLL over all contiguous width-8 windows in one context. Both scores specified here require likelihood access and differ from top-1-only SWMM. Each bootstrap resample

Table 7: Defense accounting parameters at $\delta = 10^{-5}$. Rows vary K and target ϵ; Releases counts generated tokens, Noise std. is the reported Gaussian noise scale, and Gaussian multiplier is its accountant parameter. Realized accountant values match targets at the reported precision.

K	Target ϵ	Releases	Noise std.	Gaussian multiplier
8	1	384	13.0422	73.7777
8	5	384	3.0948	17.5070
8	10	384	1.7331	9.8037
16	1	768	9.2222	104.3374
16	5	768	2.1884	24.7586
16	10	768	1.2255	13.8646

treats one candidate as a single unit, keeping its member and control contexts and all random seeds together. The reduction from the Raw baseline to the DP condition is then computed using the same resampled candidates in both conditions.

For every context, we evaluate utility using the same fixed set of 16 independent valid Formal L1 strings and 16 corresponding invalid strings, where each invalid string is created by changing one element of its valid version. We compute held-out NLL from the true-token losses across these strings. Grammar accuracy is the percentage of valid/invalid pairs for which the model assigns a lower average NLL, and therefore a higher probability, to the valid string.

Table 8: Supplementary single-context width-8 WBC-window defense results. Brackets give 95% candidate-cluster bootstrap CIs. Raw−DP is the paired decrease in attack AUC (larger means less detected leakage). Grammar acc. is discrimination accuracy; Loss (pp) is its decrease from Raw in percentage points, not token NLL. Held-out NLL is in nats/token (lower is better). Utility columns have no CIs.

K	Condition	MIA AUC [95% CI]	Raw−DP [95% CI]	Grammar acc.	Loss (pp)	Held-out NLL ↓
8	Raw	0.9999 [0.9998,1.0000]	–	100.0%	–	0.703
8	DP $\epsilon = 1$	0.5001 [0.4998,0.5006]	0.4998 [0.4993,0.5001]	93.4%	6.6	2.050
8	DP $\epsilon = 5$	0.5012 [0.4996,0.5035]	0.4987 [0.4963,0.5003]	93.8%	6.2	2.037
8	DP $\epsilon = 10$	0.4997 [0.4978,0.5013]	0.5002 [0.4986,0.5021]	94.8%	5.2	2.027
16	Raw	0.9994 [0.9985,0.9999]	–	100.0%	–	0.658
16	DP $\epsilon = 1$	0.5001 [0.4998,0.5007]	0.4993 [0.4982,0.4999]	92.4%	7.6	2.078
16	DP $\epsilon = 5$	0.5003 [0.4990,0.5019]	0.4991 [0.4974,0.5005]	93.7%	6.3	2.059
16	DP $\epsilon = 10$	0.4993 [0.4970,0.5011]	0.5001 [0.4981,0.5025]	94.4%	5.6	2.027

GlobalMean and WBC-window show the same qualitative pattern: every tested DP-synthetic condition yields AUC near 0.5. The estimates provide no clear evidence of ranking separability for these evaluated attacks under the reported intervals. This result does not validate the accountant, establish robustness to other attack families, or show a favorable general-utility trade-off.

F Matched ICL–Full-SFT Comparison

The principal paradigm comparison uses Mistral-7B-Instruct-v0.2 and five frozen seeds. It crosses $K \in \{8, 16, 32\}$ with first, middle, and last anchor positions and four audit candidates per cell, yielding 180 D^+/D^- pairs per method. Each candidate has a 64-token prefix and the same 64 teacher-forced target tokens. The member uses a fixed paraphrase, whereas the control uses the embedding-nearest same-topic example from another document. For ICL, these examples enter the active context of a frozen model; for Full-SFT, they enter the corresponding parameter-updating dataset. Both worlds score the same original target. Thus this comparison concerns paraphrased target-associated evidence and cannot be equated with the exact-exposure attack.

Full-SFT updates all 7,241,732,096 parameters in BF16 with AdamW, a learning rate of 10^{-5}, global batch size two, one unshuffled pass, gradient checkpointing, DeepSpeed ZeRO-3, and CPU optimizer offload. Each of the 360 training worlds starts from the same audited initialization. Audits confirm the initialization, parameter changes, token alignment, finite losses, and causal positions for all 180 ICL and 180 Full-SFT pairs.

Table 9: Matched ICL–Full-SFT attack results. Rows report each method and ICL-minus-Full-SFT; columns give pooled AUC, its 95% seed-cluster bootstrap CI, and TPR at 1% FPR. Higher AUC/TPR means greater leakage, not better task utility. A dash marks an unreported comparison.

Method	Pooled AUC	95% CI	TPR@1%FPR
ICL	0.869	[0.790,0.952]	0.317
Full-SFT	0.688	[0.574,0.834]	0.028
Difference	0.181	[0.080,0.319]	–

Table 10: Member token geometry in the matched comparison, using the definitions in Appendix A. Columns give method estimates and ICL-minus-SFT differences with 95% CIs. Larger positive-token ratio means broader directional coverage; larger top-10% absolute share, lower entropy, and lower effective support mean more concentrated magnitude. These metrics need not rank the methods in the same direction.

Metric	ICL	Full-SFT	ICL−SFT [95% CI]
Positive-token ratio	0.612	0.581	0.031 [0.009,0.060]
Top-10% absolute share	0.585	0.491	0.094 [0.075,0.112]
Normalized entropy	0.733	0.800	−0.067 [−0.078,−0.053]
Effective-support fraction	0.335	0.441	−0.106 [−0.122,−0.085]

Stratified ICL/Full-SFT AUCs are 0.846/0.623 at $K = 8$, 0.878/0.734 at $K = 16$, and 0.895/0.710 at $K = 32$. By anchor position, they are 0.906/0.458 for first, 0.890/0.716 for middle, and 0.824/0.931 for last. Each stratum contains only 60 controls and reuses candidates. The last-position reversal therefore does not establish a general recency law. The comparison supports a difference only under this matched exposure protocol. The positive AUC-difference interval supports higher pooled ICL separability, while the geometry table distinguishes more positive positions from greater concentration of absolute signal.

G Statistical and Reproducibility Checklist

Table 11: Analysis units and safeguards across experiments. Rows identify stages, the middle column gives the resampling unit, and the final column records pairing and selection controls. Keeping repeated measurements within their target, candidate, or seed cluster avoids treating them as independent.

Stage	Resampling unit	Safeguard
Three-factor factorial experiment	complete target	form contrasts within target/draw; retain all factorial conditions; 2,000 bootstrap and 20,000 sign flips
Token distribution	complete candidate	average context seeds within candidate before profiles; exclude $K = 0$ from membership trends
Top-1 attack	complete candidate	disjoint 100/400 development/evaluation split; freeze $w = 2$; retain all K, seeds, and worlds
Defense pilot	complete candidate	pair Raw/DP replicates; exclude incomplete $K = 32$; do not attach CIs to fixed utility pairs
ICL–Full-SFT	seed cluster	preserve matched candidate/configuration units; treat stratified low-FPR values as coarse

The natural-language experiment manifest and score table each contain 36,320 rows. Their recorded SHA-256 identifiers are `4cb7c3442ddc9a084db23e46b43f84cb3c2cd2a9da2289360e825e7f4a72777a`, `23a05860c46ca9efbdd5a1e26fcaa23706301489b54512c86c52039dec740c21`, respectively. These identifiers record the exact files used in the analysis.

The three-dataset figures and tables use metrics defined in advance, reported both overall and separately for each random seed, together with the token-level profiles. For the top-1 attack, the method setting is chosen using the development examples and then fixed before evaluation on the separate held-out examples. The defense experiment contains 4,800 complete observations for $K = 8$ and $K = 16$ and excludes the incomplete $K = 32$ experiment. For the Full-SFT comparison, we

verify that all runs start from the same model initialization, that fine-tuning changes the model parameters, that the target tokens are aligned correctly, and that all member/control evaluations produce valid finite loss values.

Every numerical result stated in the main text is supported by a main table, a main figure, or this appendix. Other exploratory analyses are not used as evidence for the claims reported in the paper.