# Watermarking for Proprietary Dataset Protection

**Anonymous Authors**[1]

## Abstract

A growing body of literature suggests that training data membership inference problems are fundamentally hard tasks in modern language modeling settings. We argue that output watermarking techniques are the right gadget to make training membership tests for generative models more tractable based on prior results showing that language models exhibit residual watermark "radioactivity" under partially watermarked training datasets. We pit this watermark-based dataset inference approach head-to-head against traditional loss-based membership inference methods and show that watermarking can achieve comparable membership detection performance under an alternate set of assumptions.

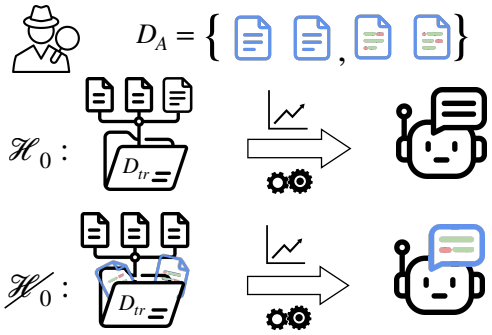


*Figure 1.* To protect a proprietary dataset from unauthorized use in training, the dataset owner (attacker) paraphrases their documents with a secret watermark key to produce $D_A$. To perform dataset inference, the attacker tests the suspect model's predictions for evidence of the watermark key. The $p$-value of the watermark test is used to conclude whether their watermarked data was included in the training dataset $D_{tr}$.

## 1. Introduction

Modern language models perform complex knowledge work of growing economic value, but the regulatory frameworks governing fair use of the web-scraped data they train on remain underdeveloped. Recent litigation suggests content owners like news websites and independent authors may be entitled to compensation for inclusion of their datasets in large-scale AI model training. Answering such data-use questions in high-stakes settings requires a concrete definition of what it means to test whether some data was included in a model's training dataset.

While the fully general training data attribution problem asks how a model's test-time behaviors are caused by specific training instances, the question at hand in contemporary fair-use deliberations is actually just *membership*. Membership inference attacks (MIAs) ask whether a specific sample was in a model's training dataset; dataset inference attacks (DIAs) generalize this to whole collections. As the more relevant setting for IP and generative-model training disputes, in this work we study the DIA setup but refer to both problem types collectively as membership problems.

**What makes training dataset membership tests challenging in modern settings?** Modern generative-model membership tests are harder than their counterparts in earlier discriminative settings (which mapped inputs to as few as 10 or 100 classes (Shokri et al., 2017)). The variable length output space matches the input cardinality, making analysis complicated, and the billion-parameter sizes of modern models make classical attribution tools like influence functions difficult to apply (Koh & Liang, 2017; Ilyas et al., 2022; Park et al., 2023). Even the definition of "a sample" is arbitrary at pretraining scale (pages, documents, sentences), and individual samples overlap heavily in n-grams, which destabilizes existing membership tests (Duan et al., 2024). Modern generative models also memorize *and* generalize along both semantic and stylistic lines, blurring the line between membership and generation attributability: sample membership is often neither necessary nor sufficient for a target output to be produced (Liu et al., 2025), and memorization rates vary widely (Cooper et al., 2025).

**Proactive Interventions for Reliable Dataset Inference.** Our work attempts to circumvent some of the difficulties described above by proactively but sparingly intervening during the construction of training datasets to make member-

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

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

ship tests easier. The essential operation will be to precisely *mark* the training dataset of a model in a way that does not significantly impact performance. This intervention will give the marker (data owner) a significant informational advantage about how the marked elements of the training dataset are expected to influence the model's test-time predictions, and most importantly, how to probe for these effects.

**Contributions:**

- We formalize the general threat model for watermark-based proactive dataset inference as a data owner-centric security game.
- We propose a randomization-based variant of the watermark detection test that ensures $p$-value validity under interactions between pretrained checkpoints, intervention samples, and specific watermark keys.
- We implement a unified experimental setting for evaluating traditional loss-based dataset inference scores and watermark-based approaches on the same paired trials, and compare the performance of each technique under its respective threat-model assumptions.
- We ablate under-explored dimensions from previous studies: per-key support vs. repetition at fixed training budget, training from scratch vs. continued pretraining, and insertion schedule shape for the marked subset.

## 2. Methodology

***Threat Model: Proactive Dataset Inference for Data Owners.*** Alice suspects that a model owner Bob will train a language model $f_{tgt}$ on $D_{tr} := D_A \cup D_{web}$, where $D_A$ is a portion Alice owns or controls and the rest is drawn from sources Alice does not control. Alice wants a score $M : f_{tgt}, \mathcal{D}_A \to (0, 1)$ that predicts whether Bob included any of $D_A$ in $D_{tr}$. Alice is allowed to modify $D_A$ using a *marking* transformation $T_m : \mathcal{X} \to \mathcal{X}_m$ before Bob accesses it. Attack success is evaluated against the ground-truth membership label $m \in \{0, 1\}$ indicating $D_A$'s inclusion in $D_{tr}$.

***Watermark-based Dataset Protection.*** Alice instantiates $T_m$ using a generative model $f_{wm} : \mathcal{X}, s \to \mathcal{X}_{wm}$ running an output watermarking scheme keyed by secret $s$, whose detector admits a test statistic $z \in \mathbb{R}$ and associated $p$-value. She marks all or part of $D_A$ via $f_{wm}$ relying on the assumption that Bob's $f_{tgt}$ trained on samples $x \in \mathcal{X}_{wm}$ will produce new generations that yield significant detection scores under $s$. A decision function $g : \mathbb{R} \to (0, 1)$ on $z$ then implements the membership score $M = g(z)$, with the $p$-value itself being the obvious choice.

**Relationship to Prior Work.** We build our study directly on two prior works. The folding construction used as our fine-tuning backbone is adapted from Hayes et al. (2025), whose paired-data MIA evaluation we extend to full-subset membership (DIA). The proactive watermarking DI approach builds on Sander et al. (2024; 2025), who show that a target model trained on watermarked samples measurably reproduces the watermark's key-specific signature. We adopt their sensitive *reading-mode detector* that tests next-token predictions (argmaxes) conditioned on watermarked prefixes rather than fully rolled-out completions. We vary per-key support fraction $1/F$ and effective epochs $E$ along the same axes, replace the detection test's parametric tail with an empirical-null randomization test (Section A.2), and benchmark against loss-based and reference-model baselines (raw loss, argmax match, min-$k\%$, zlib, rMIA-simple, rMIA, LiRA) on the same paired trials (Section C.6). **Our goal is to unify these settings and methods to calibrate claims about whether watermark-based dataset membership testing is ready for practical use.**

## 3. Experiments

Our experiments are built around a controlled data-folding design that allows us to mimic prior studies on membership and dataset inference in language models while simultaneously benchmarking the proactive watermarking approach. In a smaller finetuning regime, we systematically vary the level at which the model is exposed to the intervention data by modulating subset size and repetition during training and then we move to a larger 10B-token training regime where we ablate the insertion schedule of the marked subset and the choice of initialization: continued pretraining versus from-scratch.

**Setup.** In our experiments, we strive for depth rather than breadth. Thus, we use a single language model `Qwen3/Qwen3-1.7B` in all experiments, either as a pretrained set of weights or as the architectural specification for from-scratch experiments. Similarly, we adopt a single dataset of $N = 1500$ natural-looking but semantically isolated documents designed to be used in controlled experiments on memorization: FictionalQA. This dataset includes webtext-like documents in generic styles like news articles and blog posts about totally fictional entities and **events**—collections of related documents about the same fictional scenario (an explanatory blog post, a news article, etc.). The event grouping is what licenses our *event-split* fold construction (Section 3.1): fold boundaries respect events so that documents about the same fictional event stay together inside a single fold. We mix our controlled intervention data into a base random subset of `allenai/dolma3_mix-150B-1025` with both sources shuffled and packed into length-4096 chunks.

We score the watermark detection statistic under two surface variants. The **aligned** surface scores each original

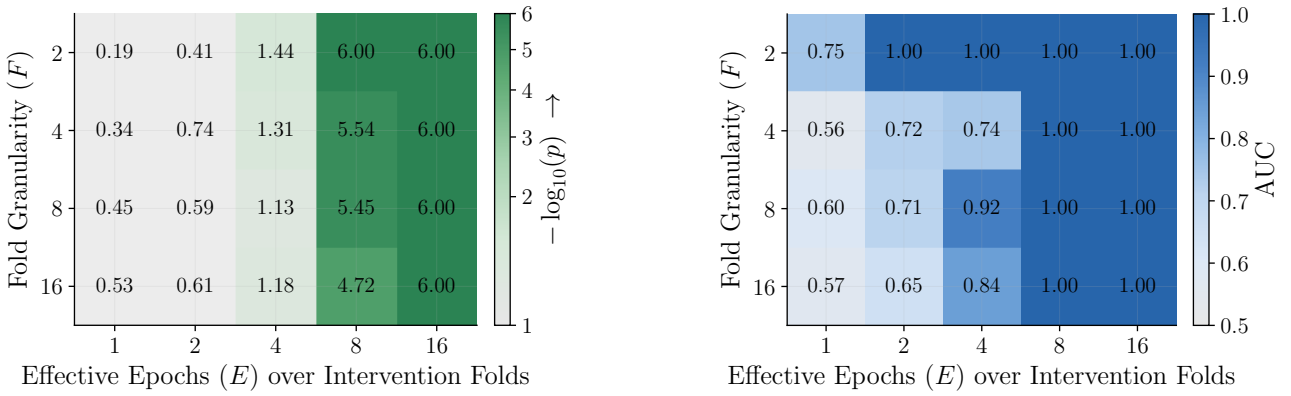


*Figure 2.* Finetuning event-split keyed signal (left) and watermark whole-model DIA AUC (right) across the $F \times E$ grid on the aligned unpacked detection surface. Scoring uses the empirical exact p-value reported as $-\log_{10} p$ and ROC-AUC is computed using matched-clean-negative models. Signal grows monotonically with $E$ and decays with $F$ as each key's per-fold footprint shrinks; the DIA view tracks the same ordering and saturates at 1.0 in the high-exposure corner. Matched clean-twin false-probe null and cross-key sham-null companions, packed-surface variants, empirical-Gaussian companions, and the SKS single-key ablation are all reported in the appendix (Sections C and D).

document in isolation and is our realistic readout. The **packed** surface scores the 4096-token chunks of multiple fictional documents packed exactly as they were at training time, and serves as an oracle baseline that quantifies how much keyed signal could be lost to train- vs. inference-time mismatches. We score the same readout under both an empirical-Gaussian and an exact empirical-null reference distribution (Section A.2), and report headline results in $-\log_{10} p$ throughout (higher means a more significant signal).

### 3.1. Finetuning: Varying Per-Key Exposure

In our finetuning-scale event-split design, multiple keyed folds of size $N/F$ are mixed into each model's training data with the total watermarked share of the training tokens held fixed across $F$. As $F$ grows, each individual key's per-fold footprint shrinks while the number of distinct keys observed by the model grows. We populate every cell of the $F \times E$ grid with enough watermarked-positive and matched clean-false-probe trials to make whole-model dataset inference comparisons stable. The per-cell training token budget is held fixed at 131M tokens. The interventional design is summarized in Table 1 (idealized per-key exposure $E/F$, shared with the SKS ablation); the matched realized normalized exposure $\hat{E}/F$ and the underlying realized $\hat{E}$ readback live in the appendix (Tables 2 and 3). In absolute terms, the watermarked share of each model's training tokens swings from a low of about 71k tokens ($\approx$0.05% of the run) at ($F$=16, $E$=1) up to about 8.3M tokens ($\approx$6.3%) at ($F$=2, $E$=16), and per-cell whole-model DIA trial counts scale with $F$, from 2/2 positive/negative trials per cell at $F$=2 to 96/96 at $F$=16 (Tables 4 and 5).

Figure 2 demonstrates the expected trends: increasing $E$ produces strongly significant keyed signal whenever the underlying support is sufficient, while increasing $F$ at fixed $E$ weakens the signal as each individual key occupies less of the corpus (Figure 10 makes this more obvious). The DIA AUC view (right panel) tracks the keyed-signal ordering and saturates at 1.0 in the high-exposure corner, while the low-exposure corner sits near chance. The matching whole-model DIA comparison against loss-based and reference-model baselines on the same paired trials is reported in Table 7 (Section C); the corresponding row-level MIA, matched false-probe-null, cross-key sham-null, and packed-surface counterparts are likewise deferred to the appendix.

### 3.2. Pretraining: Sensitivity at Scale

To address whether the keyed readout remains detectable under much heavier dilution, we conduct a second narrower batch of experiments at a 10B-token total budget. We reuse a similar $F = 2$ fold/key scaffold from finetuning (two groups of randomly sampled fictional documents of size $N/2$, not exactly the event-split construction described above) and sweep ten different insertion schedules for these two keyed folds. The two initialization regimes are continued pretraining (CPT) from a Qwen3-1.7B checkpoint and from-scratch (random init) at the same total token budget. The four single-burst schedules `(S1,E1)`, `(S2,E1)`, `(S3,E1)`, `(S4,E1)` insert the watermarked fold once at one of four step locations within the run (target $E = 1$). The remaining six schedules pair a uniform-spacing variant (denoted `U`), which uses a constant sampling rate targeting $\sim E$ epochs by end of run, with a periodic-cluster variant (denoted `P`) that schedules $E$ dumps of the entire fold at evenly spaced steps, very similar to Sander et al. (2025)'s setup. We cross

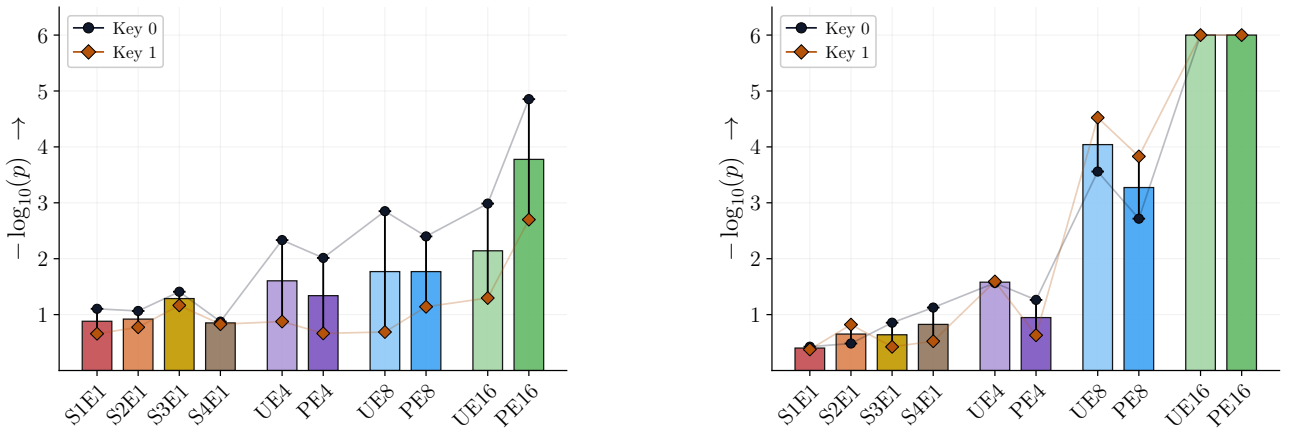


*Figure 3.* Pretraining keyed signal across the ten-schedule sweep on the aligned unpacked detection surface, scored as $-\log_{10} p$ under the exact empirical-null reference, for the CPT initialization (left) and from-scratch initialization (right). Schedule shape clearly modulates the keyed readout: periodic clusters generally separate more cleanly than the matched uniform variants at $\hat{E} \geq 8$ under CPT, while from-scratch recovers substantially stronger keyed signal at high exposure. One of the two inherited keys runs visibly warmer than the other in both panels. Matched clean-twin false-probe-null companions, packed-surface variants, watermark whole-model DIA bar charts, and loss-based / reference-model baselines are all reported in the appendix (Section E).

these with $E \in \{4, 8, 16\}$. The idealized exposure profile of these schedules at $F = 2$ is summarized in Table 13; the matched realized normalized exposure tables and per-schedule realized $\hat{E}$ values for both initialization regimes live in the appendix (Tables 16 to 19). Main body figures use the exact empirical $p$-value for both CPT and from-scratch training. Each schedule contributes 2 watermarked-positive and 2 matched clean-negative whole-model trials per init, and (mirroring Sander et al. (2025)) the watermarked-token share of each 10.49B-token training run is small in absolute terms—about $0.5M$ tokens ($\approx 0.005\%$) for the four single-burst $E$=1 schedules, rising to about $8M$ tokens ($\approx 0.077\%$) at the $E$=16 schedules (Tables 20 to 23).

**Schedule and Initialization Effects.** Across both inits, schedule shape matters: the four single-burst $E$=1 schedules are not distinguishably higher than their null counterparts in Figure 22, while the multi-insertion $U$ and $P$ schedules become clearly separable from null as $\hat{E}$ grows. Under aligned testing the $P$ and $U$ schedules are similar, but the more sensitive packed-surface readings (Figures 24 and 25) show that the $P$ setting produces a stronger signal for CPT. From-scratch recovers substantially stronger keyed signal than CPT at high $\hat{E}$, while at low $\hat{E}$ the two inits are comparable, though Figure 28 shows slightly more separability at $\hat{E} = 4$ for from-scratch than for CPT. Figures 30 and 31 also show that the $U$ sampling strategy produces more signal in the from-scratch setting.

**Fold/Key Asymmetry.** The pretraining experiments reuse the same two folds and keys that the finetuning SKS $F = 2$ row was conducted with, and at the smaller finetuning scale one of those two keys already produces systematically larger detection statistics than the other under matched-clean conditions, even when no watermark training signal is in play (right side of Figures 13 and 21). This handedness is visible again in Figure 3 where Key 0 consistently produces higher signal than the other under matched conditions (more obvious in Figure 24). This suggests that per-key interactions with the data and/or the pretrained checkpoint can produce scenarios where false-positive rates exceed what the nominal $p$-value threshold would indicate. It is possible that prior studies filtered for well-behaved watermark keys, and our systematic evaluation just reports this issue more transparently.

## 4. Conclusion

We find the watermark-based dataset inference approach to be a promising alternative to traditional loss-based DIA methods. Despite our randomization-based detection test yielding a well-calibrated empirical null, the possibility for elevated single-key readings in any setting presents a previously overlooked deployment challenge. Further, the sensitive reading-mode detection regime from prior work may itself be impractical depending on the API access the data owner has against the suspect target model. In head-to-head comparison, the loss-based baselines are essentially the more performant DIA scorers in raw AUC terms. However, they assume greater access to the target model, hidden calibration samples, or reference models. Which approach is more useful in practice therefore comes down to which set of assumptions is more realistic for the data owner.

## References


Cooper, A. F., Gokaslan, A., Ahmed, A., Cyphert, A. B., De Sa, C., Lemley, M. A., Ho, D. E., and Liang, P. Extracting memorized pieces of (copyrighted) books from open-weight language models. *arXiv preprint arXiv:2505.12546*, 2025.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024.

Fernandez, P., Sander, T., Elsahar, H., Chang, H., Souček, T., Lacatusu, V., Tran, T., Rebuffi, S.-A., and Mourachko, A. How good is post-hoc watermarking with language model rephrasing? 2025.

Hayes, J., Shumailov, I., Choquette-Choo, C. A., Jagielski, M., Kaissis, G., Nasr, M., Ghalebikesabi, S., Annamalai, M. S. M. S., Mireshghallah, N., Shilov, I., et al. Exploring the limits of strong membership inference attacks on large language models. *arXiv preprint arXiv:2505.18773*, 2025.

Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., and Madry, A. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022. URL https://arxiv.org/abs/2202.00622.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017. URL https://arxiv.org/abs/1703.04730.

Liu, K. Z., Choquette-Choo, C. A., Jagielski, M., Kairouz, P., Koyejo, S., Liang, P., and Papernot, N. Language models may verbatim complete text they were not explicitly trained on. *arXiv preprint arXiv:2503.17514*, 2025.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023. URL https://arxiv.org/abs/2303.14186.

Sander, T., Fernandez, P., Durmus, A., Douze, M., and Furon, T. Watermarking makes language models radioactive. *Advances in Neural Information Processing Systems*, 37:21079–21113, 2024. URL https://arxiv.org/abs/2402.14904.

Sander, T., Fernandez, P., Mahloujifar, S., Durmus, A., and Guo, C. Detecting benchmark contamination through watermarking. *arXiv preprint arXiv:2502.17259*, 2025. URL https://arxiv.org/abs/2502.17259.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

# A. Extended Methodological Details

**Folded Data Design.** We instantiate our folded design under two complementary support regimes that share the same $(F, E)$ axes. The primary, more realistic regime is the *event-split* setting: multiple keyed folds are mixed into each model's training data with event-level groupings preserved within each fold and with the total watermarked share of the training tokens held fixed across $F$. As $F$ grows, each individual key's per-fold footprint shrinks while the number of distinct keys present grows, simulating a deployment scenario where multiple data owners protect their own subsets with distinct keys inside a single training corpus. The simpler ablation regime is the *simple per-key support* (SKS) setting: each watermarked model trains on exactly one watermarked fold, which isolates the watermarked-fraction-versus-repetition question without sibling-key interference but is unrealistic in that no real deployed corpus would protect only one fold of one data owner's content under a single key. We treat the SKS setting as a structurally clean warm-up that anchors the more realistic event-split readout.

## A.1. Target Model Assumptions

We assume that the target model's tokenizer is known and matches the paraphraser's. Prior work shows that this assumption can be relaxed and watermark detectability still preserved through common-token filtering (Sander et al., 2024; 2025), so we omit experiments along this axis here. We also assume that the attacker can query the target model for next-token-prediction probabilities conditioned on a prefix—the "reading mode" detection regime. This is a somewhat unrealistic and costly assumption in practice if one is making calls to a production API for every token. However, the increased sensitivity of the data-forced reading-mode test is necessary at the support fractions we consider: naively-rolled-out completions from a target model that has only been mildly exposed to the watermark signature do not yield low enough $p$-values for reliable detection. Improvements in detection that would let the attacker relax the reading-mode assumption are an obvious axis of future work but are out of scope here.

## A.2. Ensuring P-value Validity

Prior watermark-DIA work reports significant $p$-values under intervention and near-null values otherwise (Sander et al., 2025), but preliminary reproduction experiments suggested that interactions between the model checkpoint, the particular samples in the intervention dataset, and the specific watermark key used can invalidate the independence assumptions on which standard parametric tail bounds depend, even under careful de-duplication and especially when the reading-mode detector is in use. Therefore, in this work we propose a standard randomization-based (permutation) hypothesis test variant that establishes a valid $p$-value based on an empirical null hypothesis under exchangeability assumptions. This method allows for configurable $p$-value precision and adds only a small computational overhead to the attacker's testing protocol; test-time cost is controlled using the vectorized PRF implementation in `textseal` (Sander et al., 2025; Fernandez et al., 2025) to evaluate many watermark secrets in parallel. We verify the validity of the resulting empirical null in practice by confirming $p$-value uniformity using a KS test on the event-split finetuning grid (Section C.7, Figure 13), with the SKS ablation showing the same property at smaller pooled-null scale (Figure 21), and fit parametric tails to extrapolate beyond the exact-null resolution under minimal additional assumptions where useful.

**Efficiency-Focused Design Choices.** The point of this folded design is to realize a large number of dataset inference trials simulating the proactive watermarking approach as well as running the loss-based methods, all while still controlling computational cost. As a result, certain independence assumptions are necessarily violated. First, the raw source data used at each $F$ level is the same fixed pool of FictionalQA documents, which makes the different experiments correlated with respect to this domain. Second, in the event-split regime each model trains on more than one fold, so a dataset inference experiment for that model with respect to fold $i$ is not fully independent of the experiment testing the same model with respect to fold $j$ which it was also trained on.

However, some of the efficiency-focused choices also help control other confounds. As we increase $E$, the fold partitions (actual selected documents) and watermark keys used remain identical, making effective epochs the only variable along that axis. Similarly, the folds that the paired clean models train on are the same underlying documents as the ones that the watermarked models train on (just before they were watermarked), controlling for natural variation in modeling difficulty between the watermarked and clean model pairs. In the event-split regime, the models are also trained on different watermarking keys at the same time, simulating a realistic aspect of the deployment scenario where different sub-datasets may be protected with different watermark keys within a single training corpus.

## B. Extended Experimental Setup Details

**Data Folding.** For each finetuning experiment, we define a fold factor $F$ and partition the FictionalQA documents into $F$ equal folds. We then train a small population of models per cell with paired clean and watermarked twins, so that each watermarked positive has a matched clean "false-probe" negative trained on the unmarked version of the same documents. To simulate the watermark-based dataset protection approach, for every fold $D_i$, we create a paraphrased version of the documents using a paraphraser model running the watermark decoding scheme with key $k_i$, producing $D_{k_i} = T_m(D_i, k_i)$. The watermarked model $f^{wm}_{\theta_i}$ is trained on a mixture that includes $D_{k_i}$ and other data, while the corresponding clean twin $f^{c}_{\theta_i}$ trains on $D_i$ in its original form before the paraphrasing transformation. We also vary the number of effective epochs $E$ over the watermarked subset within a run while holding the per-cell training-token budget fixed across $(F, E)$; samples from the base data source are not repeated. Table 1 summarizes the resulting effective per-key exposure across the grid; the event-split and SKS regimes share this idealized table.

**Metrics.** During each dataset inference simulation trial, for the watermarking methods, for positive cases we report the pooled, deduplicated reading-mode test statistic and corresponding $p$-value measured on each of the watermarked folds for each of the models that trained on it. For the negative cases, we compute the same score using all the same watermarked folds as probe data, but on the corresponding clean (twin) model trained on the unmarked version of the data. For the non-watermarking methods, the original clean samples from the fold are fed through the corresponding clean models to compute losses, and then each technique's corresponding algorithm is used to produce the row-level MIA or whole-model DIA score, potentially using the rest of the clean models as reference models. Throughout, we present results in terms of $-\log_{10} p$ for keyed signal strength, switching between an empirical-Gaussian and an exact empirical-null reference distribution where useful (Section A.2).

## C. Finetuning Event-Split Exhaustive Readout

This appendix carries the full per-cell readout of the event-split finetuning grid that the main body's Figure 2 compresses into a single combined headline panel. The main body uses the empirical-exact reference for both the keyed-signal and DIA heatmaps; the subsections below add the empirical-Gaussian companion of each, the packed-surface counterparts of the keyed/null pair, the cross-key sham-null heatmaps, the underlying watermark whole-model DIA numeric table, the realized exposure and $\hat{E}$ readbacks, the row-level MIA baseline table, and the null-validity panel.

*Table 1.* Idealized per-key exposure $E/F$ for the finetuning grid, expressed as a multiple of the per-key support that a single epoch over a single fold would produce. Distinct watermarked tokens scale as $1/F$; epoch repetition $E$ multiplies how many times those tokens are seen during training. The event-split and SKS regimes share this idealized table.

| | $E = 1$ | $E = 2$ | $E = 4$ | $E = 8$ | $E = 16$ |
|---|---|---|---|---|---|
| $F = 2$ | 0.5 | 1.0 | 2.0 | 4.0 | 8.0 |
| $F = 4$ | 0.25 | 0.5 | 1.0 | 2.0 | 4.0 |
| $F = 8$ | 0.125 | 0.25 | 0.5 | 1.0 | 2.0 |
| $F = 16$ | 0.0625 | 0.125 | 0.25 | 0.5 | 1.0 |

### C.1. Empirical-Gaussian and Packed-Surface Heatmap Companions

Figures 4 to 7 carry the four event-split keyed/null pairs across the (aligned, packed) surface and (exact, Gaussian) $p$-value-type combinations; the aligned-exact keyed half is also shown as the left panel of the main-body Figure 2, and is repeated here paired with its matched clean-twin false-probe null companion. Figures 8 and 9 carry the matching watermark whole-model DIA AUC pairs, with aligned and packed surfaces shown side-by-side under each $p$-value type.

### C.2. Event-Split Exposure Trend Curves

Figure 10 is a re-visualization of the same keyed-signal data shown in the left panel of Figure 2, re-cast on a continuous exposure axis with separate lines per $F$ to make the exposure trend and the cross-$F$ separation easier to read at a glance: at low exposure the per-key support is too small for keyed signal to lift off in any of the $F$ rows, while at high exposure the curves separate cleanly with $F$.

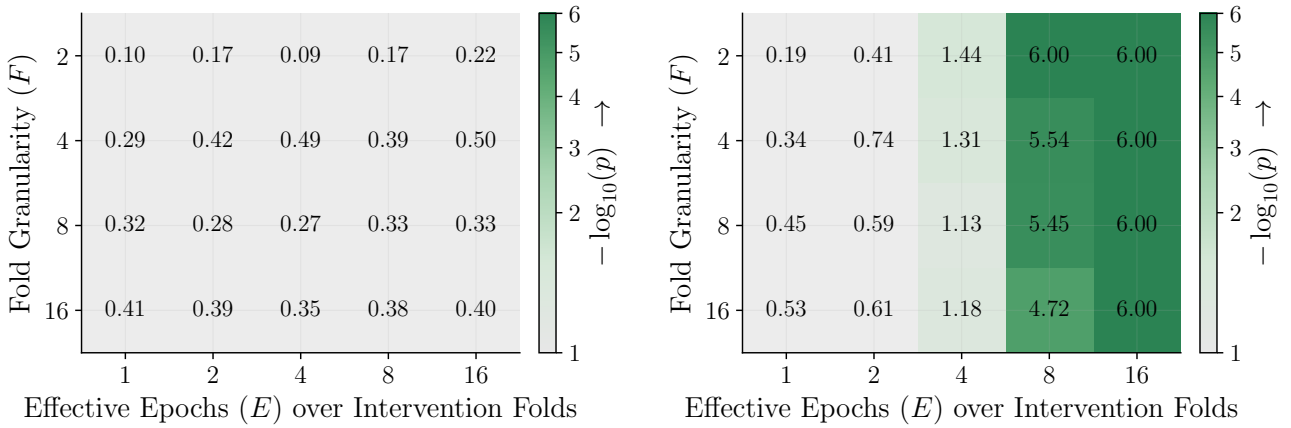


*Figure 4.* Finetuning event-split matched clean-twin false-probe null (left) and keyed signal (right) on the aligned unpacked detection surface, scored as $-\log_{10} p$ under an empirical-exact reference. The keyed half is the same surface as the left panel of Figure 2, paired here with its false-probe-null companion to validate the matched clean-twin negative as the correct baseline.

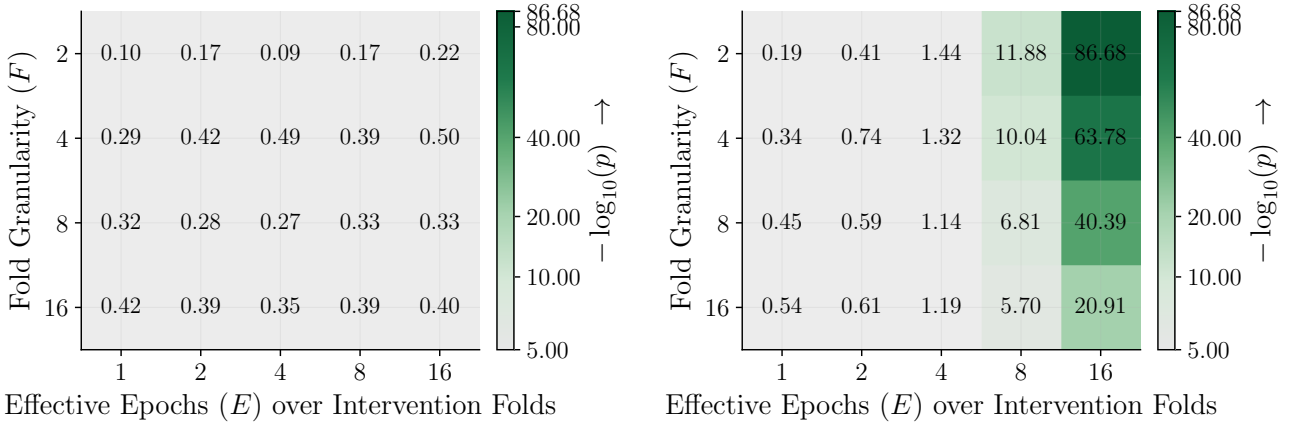


*Figure 5.* Finetuning event-split matched clean-twin false-probe null (left) and keyed signal (right) across the $F \times E$ grid on the aligned unpacked detection surface, scored as $-\log_{10} p$ under the empirical-Gaussian reference.

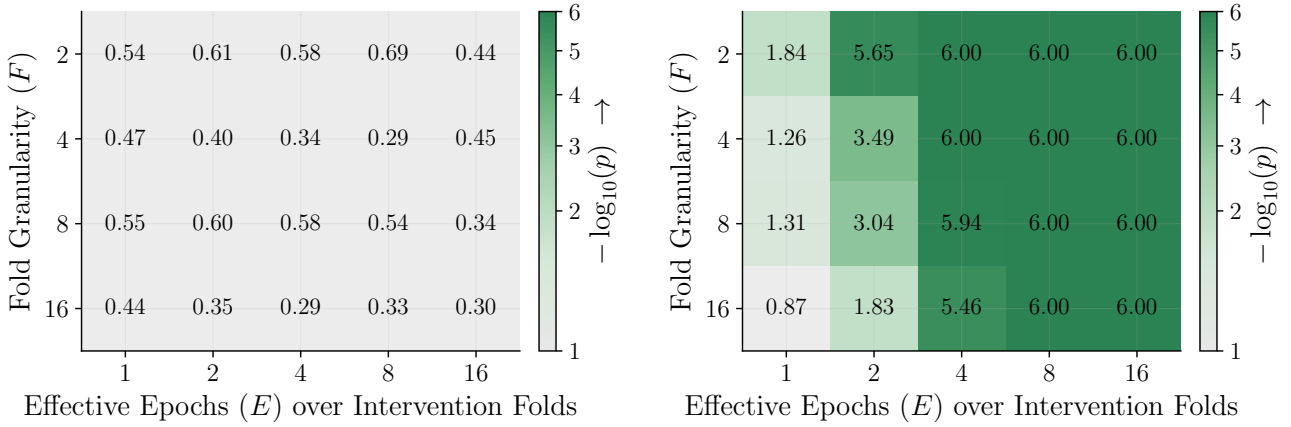


*Figure 6.* Finetuning event-split matched clean-twin false-probe null (left) and keyed signal (right) across the $F \times E$ grid on the packed detection surface, scored as $-\log_{10} p$ under the empirical-exact null. The packed surface is a more permissive oracle baseline than the aligned surface.

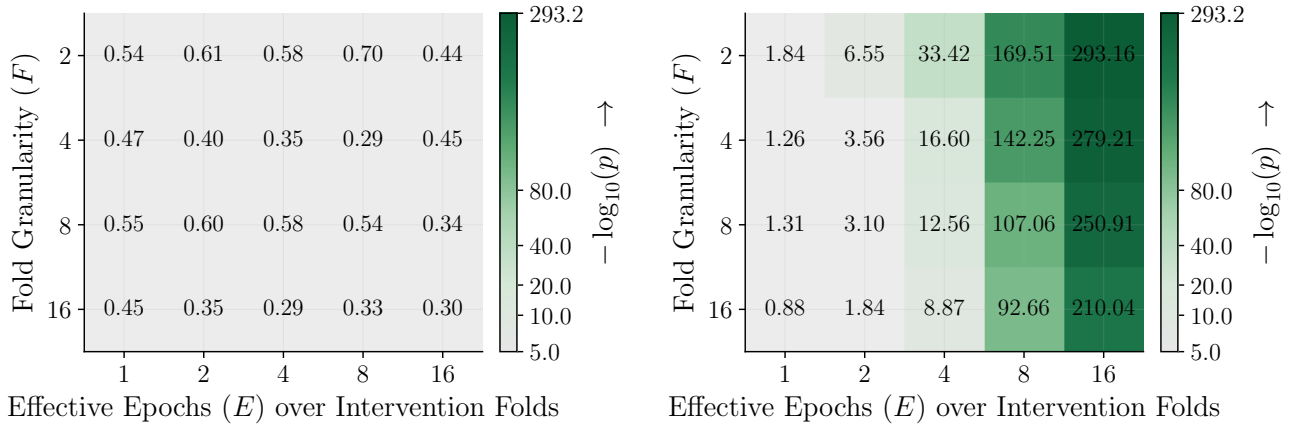


*Figure 7.* Finetuning event-split matched clean-twin false-probe null (left) and keyed signal (right) across the $F \times E$ grid on the packed detection surface, scored as $-\log_{10} p$ under the empirical-Gaussian reference.

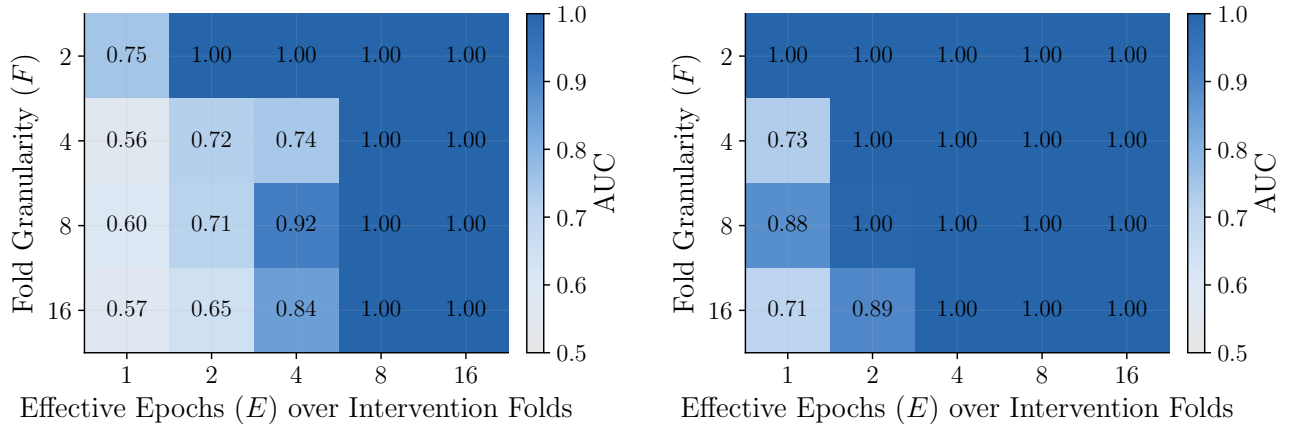


*Figure 8.* Finetuning event-split watermark whole-model DIA AUC across the $F \times E$ grid, on the aligned (left) and packed (right) detection surfaces, both scored against an empirical-exact null. The aligned half is the same surface as the right panel of Figure 2, paired here with the packed-surface companion. On the aligned surface the AUC saturates in the high-exposure corner; on the packed surface the more permissive oracle recovers several lower-exposure cells.

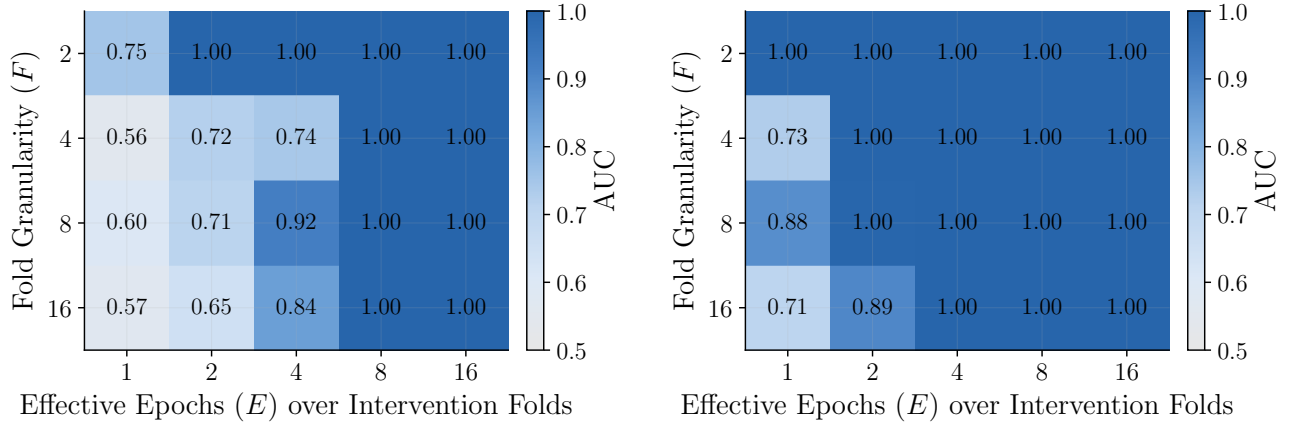


*Figure 9.* Finetuning event-split watermark whole-model DIA AUC across the $F \times E$ grid, on the aligned (left) and packed (right) detection surfaces, both scored against an empirical-Gaussian null.

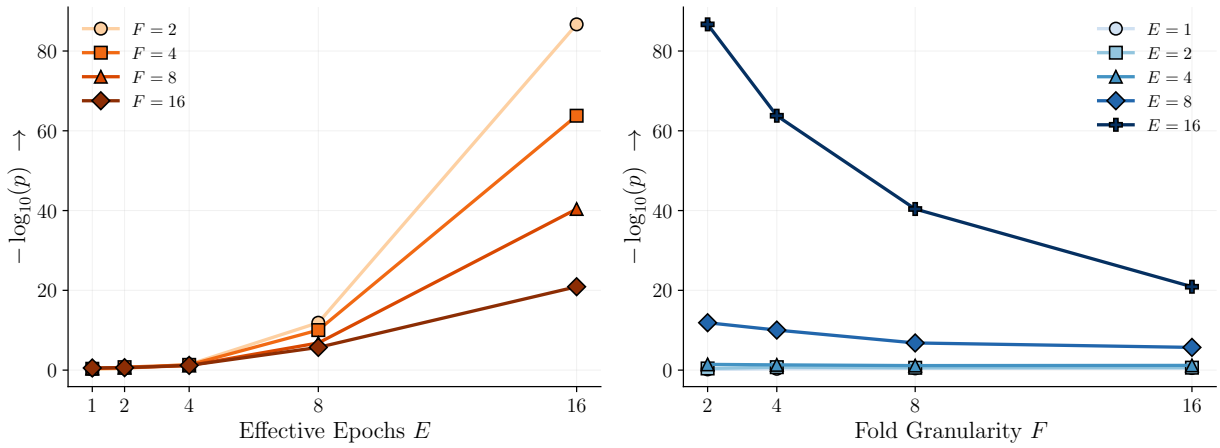


*Figure 10.* Finetuning event-split keyed-signal exposure response across the grid, tracing $-\log_{10} p$ as a function of effective per-key exposure with separate lines for each fold count $F$. This figure plots the same per-cell keyed-signal values as the left panel of Figure 2, just re-cast on a continuous exposure axis to make the trend clearer.

### C.3. Cross-Key Sham-Null Heatmaps

Figures 11 and 12 carry the cross-key sham-null heatmaps for the event-split grid, where the watermark detector is queried with a key the target model never saw in training. These act as an additional negative control beyond the matched clean-twin false-probe null, isolating per-key idiosyncrasy of the detection surface from the matched-pair design.

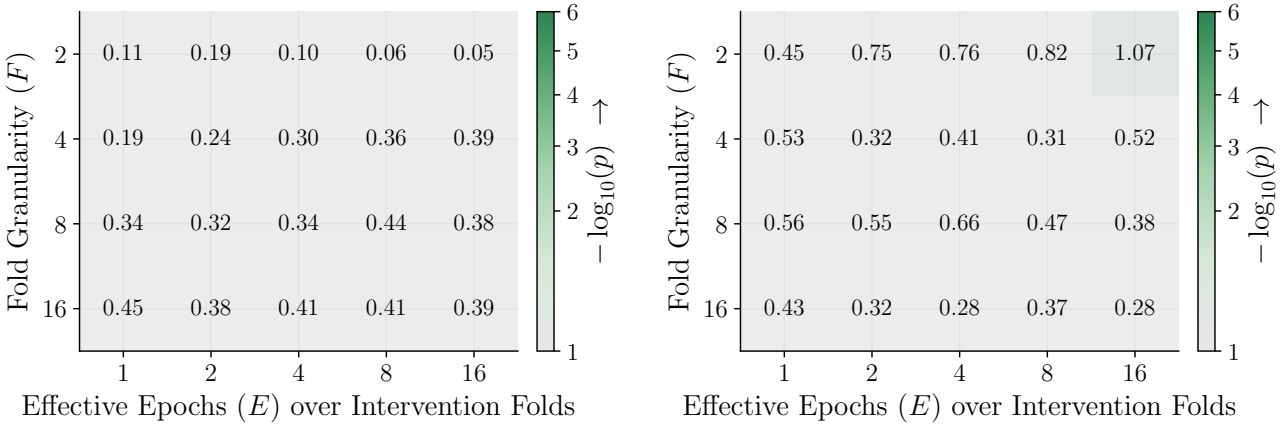


*Figure 11.* Finetuning event-split cross-key sham-null heatmap across the $F \times E$ grid, on the aligned (left) and packed (right) detection surfaces, scored as $-\log_{10} p$ under the empirical-exact null. The watermark detector is queried with a key the target model never saw in training, providing a negative control beyond the matched clean-twin false-probe null.

### C.4. Event-Split Realized Exposure and $\hat{E}$ Per-Cell Readbacks

Table 2 reports the realized normalized exposure $\hat{E}/F$ per cell of the event-split finetuning grid. Table 3 carries the underlying realized $\hat{E}$ readback. The corresponding idealized epoch counts $E$ are the column headers of Table 1, which the event-split construction shares with SKS.

### C.5. Event-Split Training Scale and Trial Geometry

Table 4 reports the per-cell watermarked-token totals (mean, min–max across paired models) and the corresponding fraction of each model's 131M-token training budget. Table 5 reports the per-cell paired-model counts and the resulting $n_+/n_-$ trial counts that drive each cell's whole-model DIA AUC; both are summarized in the in-text pointer in Section 3.1.

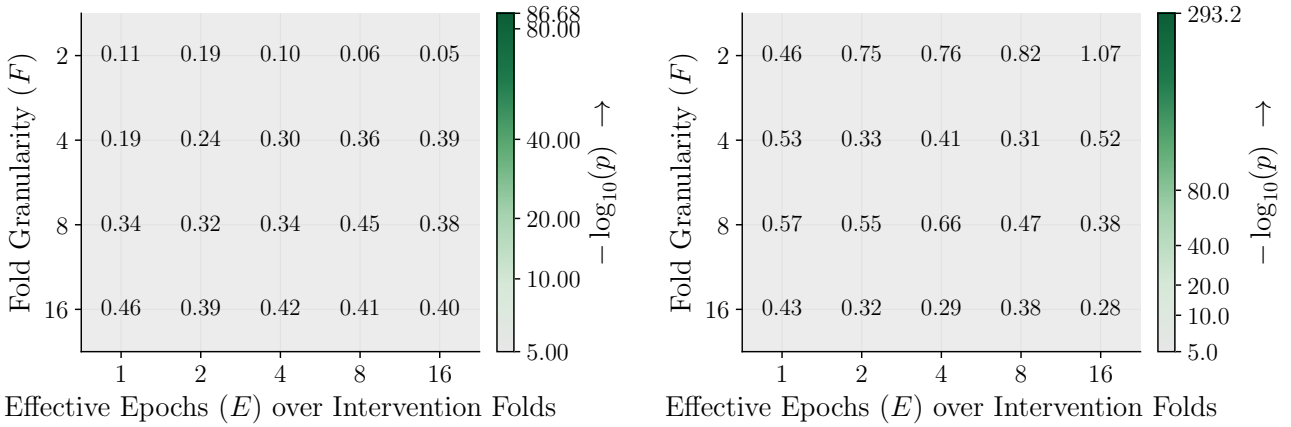


*Figure 12.* Finetuning event-split cross-key sham-null heatmap across the $F \times E$ grid, on the aligned (left) and packed (right) detection surfaces, scored as $-\log_{10} p$ under the empirical-Gaussian null.

*Table 2.* Event-split finetuning: realized normalized exposure summary ($\hat{E}/F$).

| | $E = 1$ | $E = 2$ | $E = 4$ | $E = 8$ | $E = 16$ |
|---|---|---|---|---|---|
| $F = 2$ | 0.5612 | 1.1428 | 2.2880 | 4.3259 | 8.2775 |
| $F = 4$ | 0.2944 | 0.5955 | 1.1603 | 2.2555 | 4.3436 |
| $F = 8$ | 0.1490 | 0.2996 | 0.5883 | 1.1481 | 2.2133 |
| $F = 16$ | 0.0837 | 0.1699 | 0.3310 | 0.6457 | 1.2488 |

*Table 3.* Event-split finetuning: realized exposure summary ($\hat{E}$).

| | $E = 1$ | $E = 2$ | $E = 4$ | $E = 8$ | $E = 16$ |
|---|---|---|---|---|---|
| $F = 2$ | 1.1223 | 2.2857 | 4.5760 | 8.6519 | 16.5550 |
| $F = 4$ | 1.1776 | 2.3818 | 4.6412 | 9.0218 | 17.3744 |
| $F = 8$ | 1.1922 | 2.3970 | 4.7062 | 9.1846 | 17.7066 |
| $F = 16$ | 1.3396 | 2.7188 | 5.2960 | 10.3307 | 19.9814 |

### C.6. Event-Split Loss-Based and Reference-Model Baselines

Tables 6 and 7 report the row-level MIA and fold-level whole-model DIA comparisons for the event-split finetuning grid, with the watermark detector's own scores reported alongside the loss-based and reference-model baselines on the same paired trials. The watermark whole-model DIA AUC values reproduce those visualized in the right panel of Figure 2.

### C.7. Event-Split Null-Validity

Figure 13 confirms that, once pooled across many distinct positive keys, the empirical exact null is close to uniform at the 1M-null scale and respects standard tail-rate thresholds, supporting the use of the empirical-null permutation test as the headline reference distribution for the keyed readout. Per-key idiosyncrasy is a separate concern visible in the right-hand trace: at very small fold counts a single key can run systematically warm even when the pooled null is well-calibrated, and the inherited $F = 2$ scaffold reused in pretraining (Section 3.2) carries one such warm key.

## D. Finetuning SKS Exhaustive Readout (Ablation)

This appendix reports the simple per-key support (SKS) ablation, where each model trains on exactly one watermarked fold so the per-model fictional support fraction shrinks as $1/F$ instead of being held fixed across $F$. SKS is structurally simpler than the event-split regime (Section C) but unrealistic in that no real deployed corpus would protect only one fold of one data owner's content under a single key. The same idealized per-key exposure (Table 1) governs both regimes; what differs is sibling support and the per-cell positive/negative trial budget (Table 11). We use SKS as a watermark-only sanity check on the per-key scaling story without sibling-key interference, and run the loss-based and reference-model comparison only

*Table 4.* Event-split finetuning: training scale context. Per-cell watermark token totals are relative to 131,072,000 train tokens per run. The percent columns report watermark-token share of total train tokens.

| Cell | Target $(E/F)$ | WM tokens mean | WM tokens min-max | Mean % | Range % |
|---|---|---|---|---|---|
| (F2,E1) | 0.5000 | 563,338 | 536,707-589,968 | 0.430% | 0.409%-0.450% |
| (F2,E2) | 1.0000 | 1,143,063 | 1,081,608-1,204,518 | 0.872% | 0.825%-0.919% |
| (F2,E4) | 2.0000 | 2,288,174 | 2,171,410-2,404,939 | 1.746% | 1.657%-1.835% |
| (F2,E8) | 4.0000 | 4,318,238 | 4,293,656-4,342,820 | 3.295% | 3.276%-3.313% |
| (F2,E16) | 8.0000 | 8,294,376 | 8,239,067-8,349,686 | 6.328% | 6.286%-6.370% |
| (F4,E1) | 0.2500 | 287,473 | 221,238-368,730 | 0.219% | 0.169%-0.281% |
| (F4,E2) | 0.5000 | 579,726 | 462,961-688,296 | 0.442% | 0.353%-0.525% |
| (F4,E4) | 1.0000 | 1,122,237 | 1,028,347-1,306,943 | 0.856% | 0.785%-0.997% |
| (F4,E8) | 2.0000 | 2,180,628 | 1,978,851-2,339,387 | 1.664% | 1.510%-1.785% |
| (F4,E16) | 4.0000 | 4,200,108 | 3,802,016-4,490,312 | 3.204% | 2.901%-3.426% |
| (F8,E1) | 0.1250 | 141,688 | 90,134-200,753 | 0.108% | 0.069%-0.153% |
| (F8,E2) | 0.2500 | 286,022 | 217,141-372,827 | 0.218% | 0.166%-0.284% |
| (F8,E4) | 0.5000 | 558,728 | 462,961-708,781 | 0.426% | 0.353%-0.541% |
| (F8,E8) | 1.0000 | 1,090,314 | 917,728-1,397,077 | 0.832% | 0.700%-1.066% |
| (F8,E16) | 2.0000 | 2,102,871 | 1,757,613-2,454,103 | 1.604% | 1.341%-1.872% |
| (F16,E1) | 0.0625 | 70,844 | 28,679-131,104 | 0.054% | 0.022%-0.100% |
| (F16,E2) | 0.1250 | 143,011 | 90,134-208,947 | 0.109% | 0.069%-0.159% |
| (F16,E4) | 0.2500 | 279,364 | 221,238-360,536 | 0.213% | 0.169%-0.275% |
| (F16,E8) | 0.5000 | 545,157 | 446,573-729,266 | 0.416% | 0.341%-0.556% |
| (F16,E16) | 1.0000 | 1,051,435 | 815,303-1,376,592 | 0.802% | 0.622%-1.050% |

*Table 5.* Event-split finetuning: model and watermark DIA trial geometry. The WM and clean model columns count target models per cell, and the WM DIA trial column counts the positive/negative pooled trials used for the watermark DIA AUC.

| Cell | WM models $n_+$ | Clean models $n_-$ | WM DIA trials $n_+/n_-$ |
|---|---|---|---|
| (F2,E1) | 2 | 2 | 2 / 2 |
| (F2,E2) | 2 | 2 | 2 / 2 |
| (F2,E4) | 2 | 2 | 2 / 2 |
| (F2,E8) | 2 | 2 | 2 / 2 |
| (F2,E16) | 2 | 2 | 2 / 2 |
| (F4,E1) | 6 | 6 | 12 / 12 |
| (F4,E2) | 6 | 6 | 12 / 12 |
| (F4,E4) | 6 | 6 | 12 / 12 |
| (F4,E8) | 6 | 6 | 12 / 12 |
| (F4,E16) | 6 | 6 | 12 / 12 |
| (F8,E1) | 12 | 12 | 48 / 48 |
| (F8,E2) | 12 | 12 | 48 / 48 |
| (F8,E4) | 12 | 12 | 48 / 48 |
| (F8,E8) | 12 | 12 | 48 / 48 |
| (F8,E16) | 12 | 12 | 48 / 48 |
| (F16,E1) | 12 | 12 | 96 / 96 |
| (F16,E2) | 12 | 12 | 96 / 96 |
| (F16,E4) | 12 | 12 | 96 / 96 |
| (F16,E8) | 12 | 12 | 96 / 96 |
| (F16,E16) | 12 | 12 | 96 / 96 |

on the more realistic event-split regime where the row-level baselines are meaningful.

*Table 6.* Event-split finetuning: row-level MIA AUC comparison. Entries marked N/A indicate statistics that are not estimable in the available cell geometry; for example, LiRA in F=2 cells lacks sufficient in-reference models.

| | Watermark Readout | Loss-based Row MIA | | | Ref-model Row MIA | | |
|---|---|---|---|---|---|---|---|
| Cell | WM $-\log_{10}(p)$ | Raw-loss | Argmax | min-$k_{10}$ | rMIA-simple | rMIA | LiRA |
| (F2,E1) | 0.1894 | 0.6096 | 0.6083 | 0.6292 | 1.0000 | 0.9993 | N/A |
| (F2,E2) | 0.4135 | 0.6988 | 0.6987 | 0.7318 | 1.0000 | 0.9993 | N/A |
| (F2,E4) | 1.4385 | 0.8521 | 0.8508 | 0.8991 | 1.0000 | 0.9993 | N/A |
| (F2,E8) | 11.8808 | 0.9744 | 0.9744 | 0.9707 | 0.9999 | 0.9993 | N/A |
| (F2,E16) | 86.6829 | 0.9832 | 0.9870 | 0.9683 | 0.9998 | 0.9993 | N/A |
| (F4,E1) | 0.3363 | 0.6041 | 0.6013 | 0.6237 | 0.9760 | 0.9805 | 0.8050 |
| (F4,E2) | 0.7380 | 0.6901 | 0.6896 | 0.7229 | 0.9985 | 0.9978 | 0.9165 |
| (F4,E4) | 1.3158 | 0.8360 | 0.8336 | 0.8813 | 1.0000 | 0.9993 | 0.9415 |
| (F4,E8) | 10.0393 | 0.9711 | 0.9691 | 0.9718 | 1.0000 | 0.9993 | 0.9416 |
| (F4,E16) | 63.7768 | 0.9745 | 0.9788 | 0.9608 | 0.9997 | 0.9992 | 0.9707 |
| (F8,E1) | 0.4487 | 0.6022 | 0.6010 | 0.6207 | 0.9645 | 0.9686 | 0.9668 |
| (F8,E2) | 0.5877 | 0.6852 | 0.6849 | 0.7163 | 0.9952 | 0.9946 | 0.9981 |
| (F8,E4) | 1.1352 | 0.8306 | 0.8284 | 0.8731 | 0.9992 | 0.9986 | 0.9996 |
| (F8,E8) | 6.8089 | 0.9564 | 0.9542 | 0.9552 | 0.9996 | 0.9990 | 0.9992 |
| (F8,E16) | 40.3926 | 0.9672 | 0.9717 | 0.9530 | 0.9995 | 0.9990 | 0.9991 |
| (F16,E1) | 0.5373 | 0.5975 | 0.5965 | 0.6157 | 0.9302 | 0.9341 | 0.9486 |
| (F16,E2) | 0.6138 | 0.6774 | 0.6758 | 0.7068 | 0.9771 | 0.9768 | 0.9911 |
| (F16,E4) | 1.1855 | 0.8148 | 0.8098 | 0.8471 | 0.9904 | 0.9897 | 0.9926 |
| (F16,E8) | 5.6955 | 0.9148 | 0.9166 | 0.9064 | 0.9903 | 0.9917 | 0.9898 |
| (F16,E16) | 20.9064 | 0.9294 | 0.9355 | 0.9132 | 0.9959 | 0.9962 | 0.9940 |

*Table 7.* Event-split finetuning: fold-level whole-model DIA AUC comparison. Entries marked N/A indicate statistics that are not estimable in the available cell geometry; for example, LiRA in F=2 cells lacks sufficient in-reference models.

| | Watermark DIA | | Loss-based DIA | | | Ref-model DIA | | |
|---|---|---|---|---|---|---|---|---|
| Cell | Aligned | Packed | Raw-loss | Argmax | min-$k_{10}$ | rMIA-simple | rMIA | LiRA |
| (F2,E1) | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | N/A |
| (F2,E2) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | N/A |
| (F2,E4) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | N/A |
| (F2,E8) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | N/A |
| (F2,E16) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | N/A |
| (F4,E1) | 0.5556 | 0.7292 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.1181 |
| (F4,E2) | 0.7222 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.4861 |
| (F4,E4) | 0.7431 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.4514 |
| (F4,E8) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.4514 |
| (F4,E16) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.8056 |
| (F8,E1) | 0.5968 | 0.8815 | 1.0000 | 0.9987 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (F8,E2) | 0.7118 | 0.9974 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (F8,E4) | 0.9162 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (F8,E8) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (F8,E16) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (F16,E1) | 0.5675 | 0.7062 | 0.9787 | 0.9702 | 0.9887 | 1.0000 | 1.0000 | 1.0000 |
| (F16,E2) | 0.6477 | 0.8949 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (F16,E4) | 0.8439 | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (F16,E8) | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (F16,E16) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

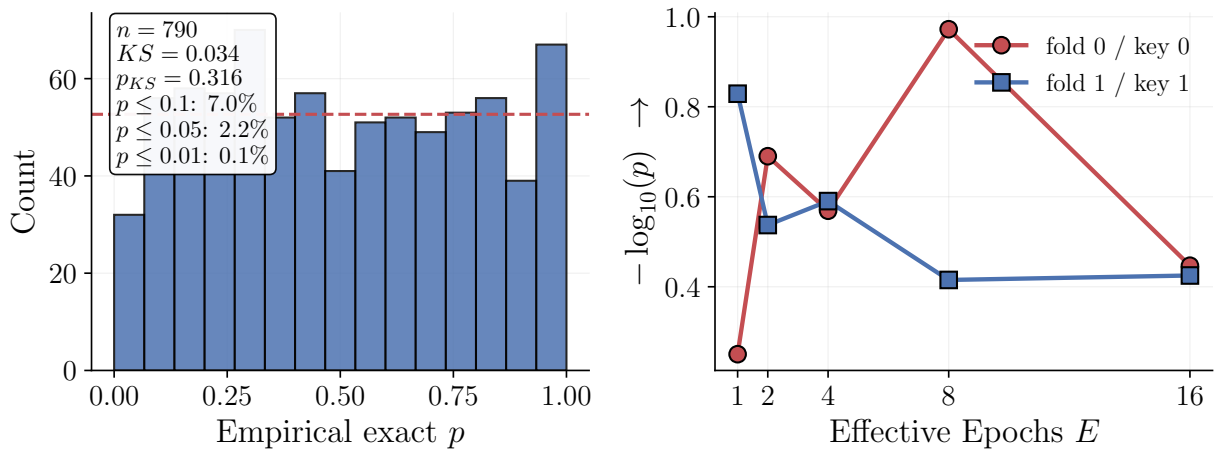


*Figure 13.* Event-grid null-validity panel. **Left:** histogram of pooled empirical exact $p$-values from the packed matched-clean-negative whole-model readings attached to packed keyed watermark surfaces across the $(F, E)$ event grid (per-target exact null scores from packed keyed-surface `whole_model_exact_dia`, converted as $p_{\text{exact}} = 10^{-\text{sig}_{\text{exact}}}$). The annotated KS statistic and $p$-value are from a one-sample Kolmogorov-Smirnov test (`scipy.stats.kstest`) of these pooled $p$-values against $\text{Uniform}(0, 1)$; the dashed horizontal line is the expected per-bin count under a uniform histogram with the plotted binning. **Right:** $-\log_{10} p$ trace of the $F = 2$ warm-key slice of the same packed null family plotted against $E$, shown for context only and not part of the KS test on the left.

### D.1. SKS Empirical-Gaussian and Packed-Surface Heatmap Companions

Figures 14 to 19 carry the SKS keyed/null pair and DIA AUC heatmap pair on the aligned-exact, aligned-Gaussian, packed-exact, and packed-Gaussian detection surfaces.

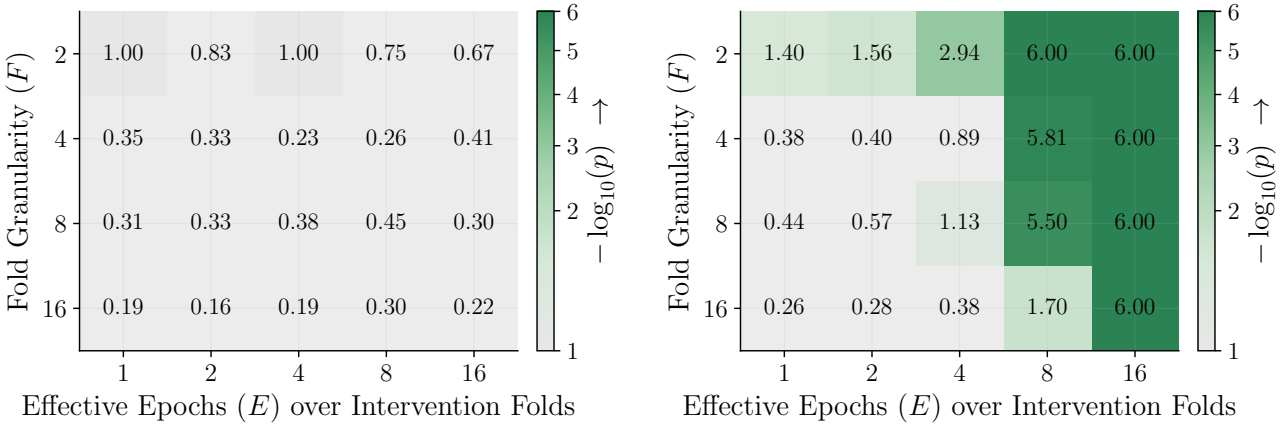


*Figure 14.* Finetuning SKS matched clean-twin false-probe null (left) and keyed signal (right) across the $F \times E$ grid on the aligned unpacked detection surface, scored as $-\log_{10} p$ under an empirical-exact reference. The false-probe null stays quiet on the same surface, validating the matched clean-twin negative as the right baseline against which to read the keyed map; the keyed signal grows monotonically with $E$ and decays with $F$ as the per-key support fraction $1/F$ shrinks.

### D.2. SKS Exposure Trend Curves

Figure 20 is a re-visualization of the same keyed-signal data shown in the right panel of Figure 14, re-cast on a continuous exposure axis with separate lines per $F$ to make the exposure trend and the cross-$F$ separation easier to read at a glance: at low exposure the per-key support is too small for keyed signal to lift off in any of the $F$ rows, while at high exposure the curves separate cleanly with $F$.

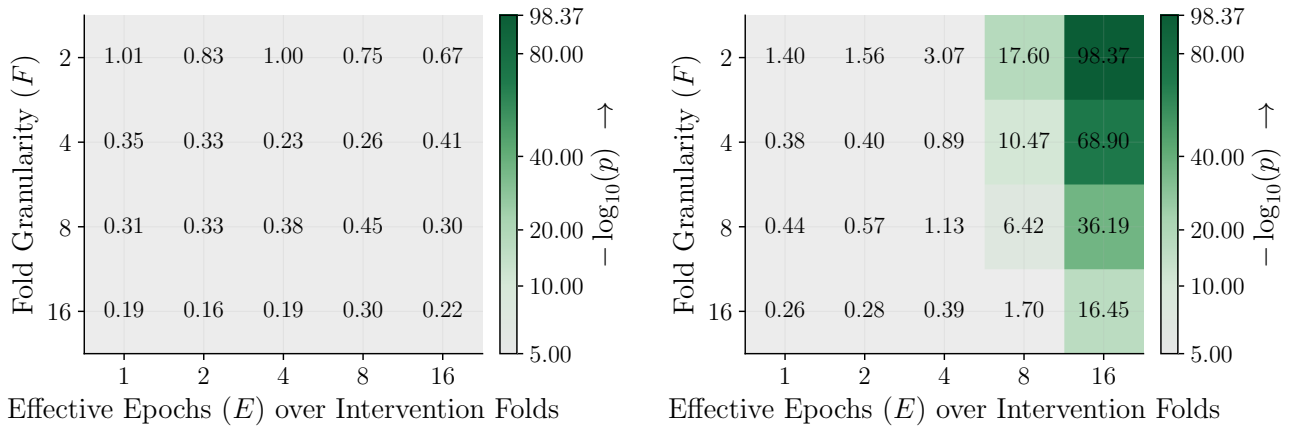


*Figure 15.* Finetuning SKS matched clean-twin false-probe null (left) and keyed signal (right) across the $F \times E$ grid on the aligned unpacked detection surface, scored as $-\log_{10} p$ under the empirical-Gaussian reference.

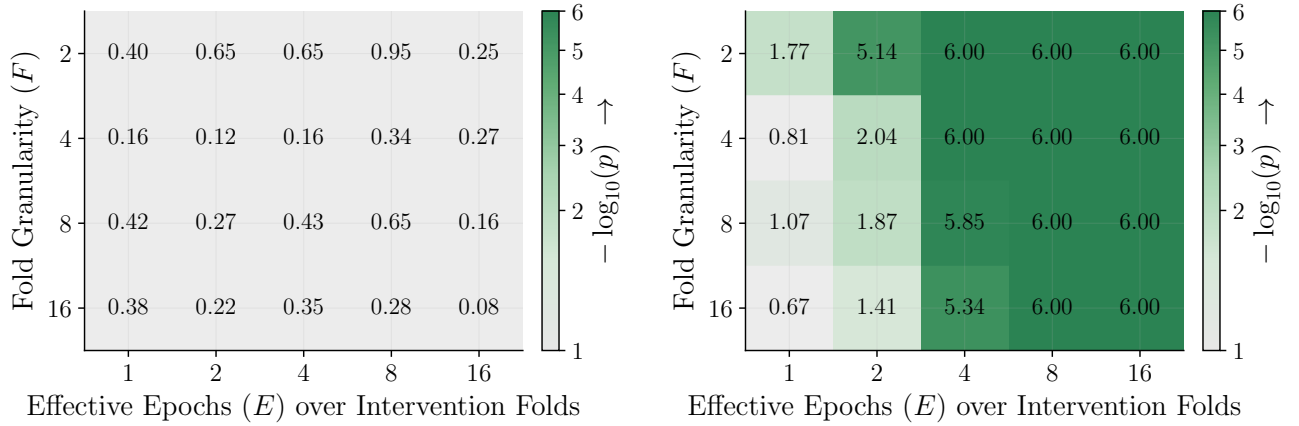


*Figure 16.* Finetuning SKS matched clean-twin false-probe null (left) and keyed signal (right) across the $F \times E$ grid on the packed detection surface, scored as $-\log_{10} p$ under the empirical-exact null. The packed surface is a more permissive oracle baseline than the aligned surface.

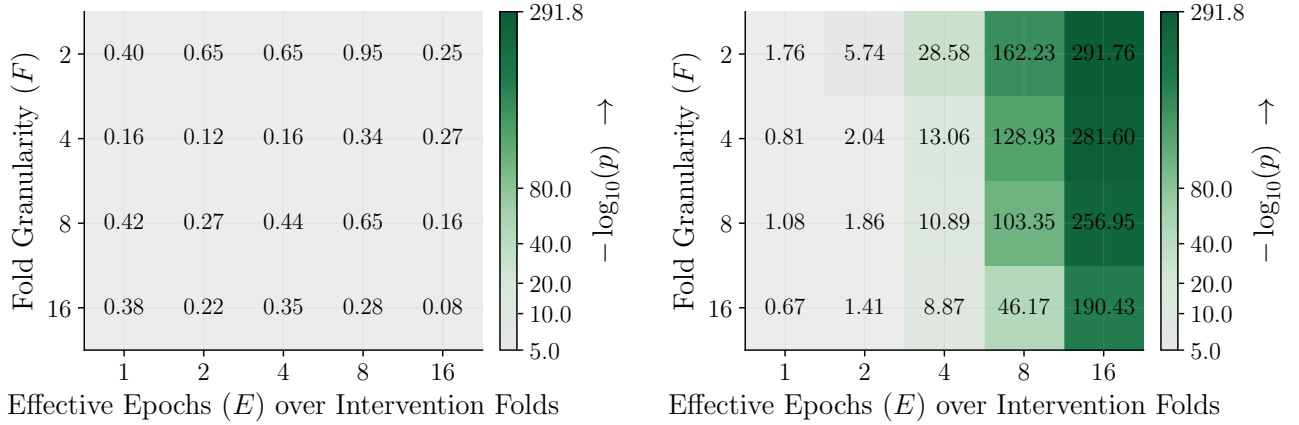


*Figure 17.* Finetuning SKS matched clean-twin false-probe null (left) and keyed signal (right) across the $F \times E$ grid on the packed detection surface, scored as $-\log_{10} p$ under the empirical-Gaussian reference.

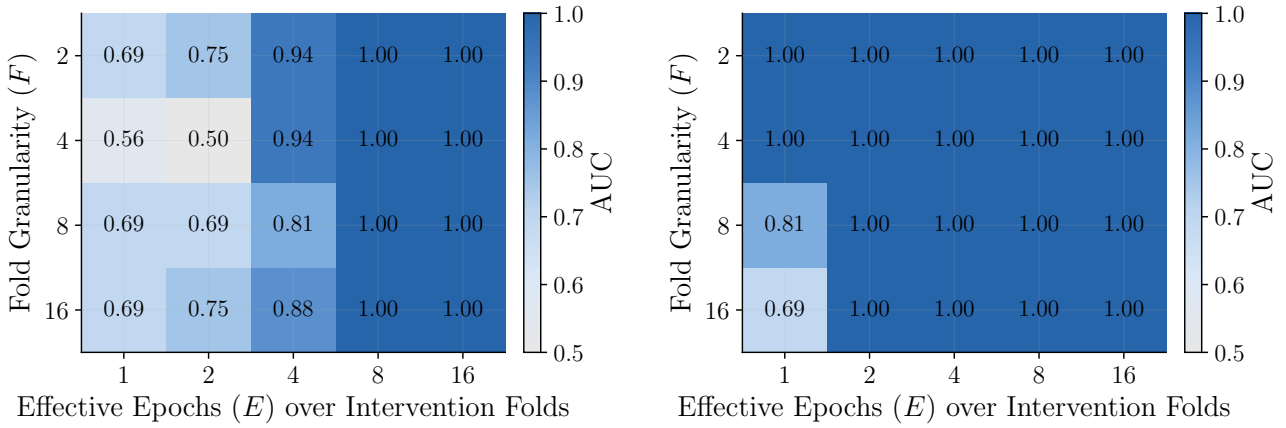


*Figure 18.* Finetuning SKS watermark whole-model DIA AUC across the $F \times E$ grid, on the aligned (left) and packed (right) detection surfaces, both scored against an empirical-exact null. On the aligned surface the AUC saturates at 1.0 from $E = 8$ onward at every $F$, while the lowest-$E$ corner remains coarse with only eight trials per cell. On the packed surface, the more permissive oracle recovers several of those low-exposure cells, reaching 1.0 one to two $E$-steps earlier across the grid.

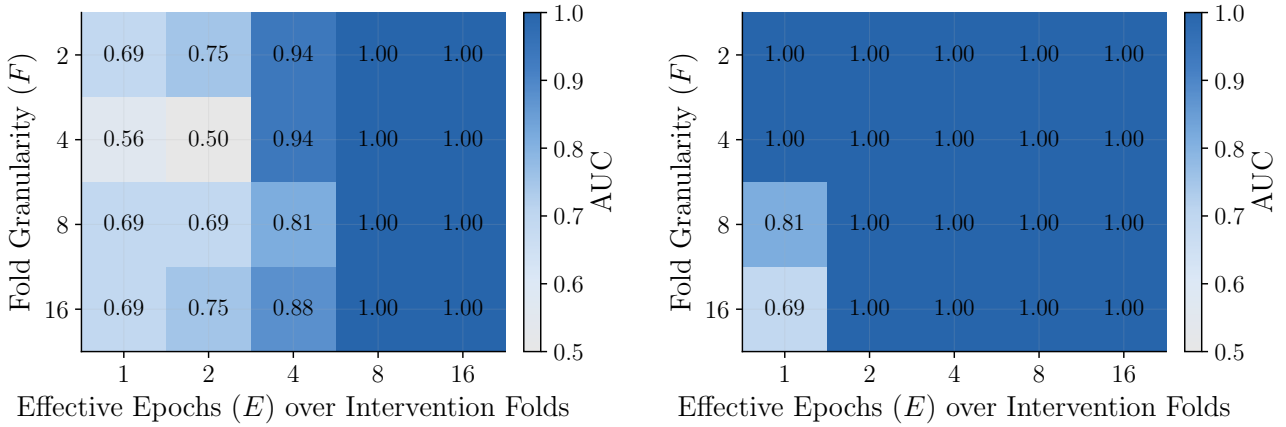


*Figure 19.* Finetuning SKS watermark whole-model DIA AUC across the $F \times E$ grid, on the aligned (left) and packed (right) detection surfaces, both scored against an empirical-Gaussian null.

### D.3. SKS Realized Exposure and $\hat{E}$ Per-Cell Readbacks

Table 8 reports the realized normalized exposure $\hat{E}/F$ per cell of the finetuning SKS grid, the direct realized counterpart to the idealized $E/F$ values in Table 1. Table 9 carries the underlying realized $\hat{E}$ readback. The discrepancy between idealized and realized values here is the realized overshoot of the watermarked subset's effective epoch count relative to the planned schedule.

*Table 8.* SKS finetuning: realized normalized exposure summary ($\hat{E}/F$).

| | $E = 1$ | $E = 2$ | $E = 4$ | $E = 8$ | $E = 16$ |
|---|---|---|---|---|---|
| $F = 2$ | 0.5557 | 1.1246 | 2.2800 | 4.3204 | 8.3766 |
| $F = 4$ | 0.2755 | 0.5584 | 1.1609 | 2.2062 | 4.3422 |
| $F = 8$ | 0.1549 | 0.2884 | 0.5772 | 1.1895 | 2.1984 |
| $F = 16$ | 0.0911 | 0.1797 | 0.3477 | 0.6667 | 1.2930 |

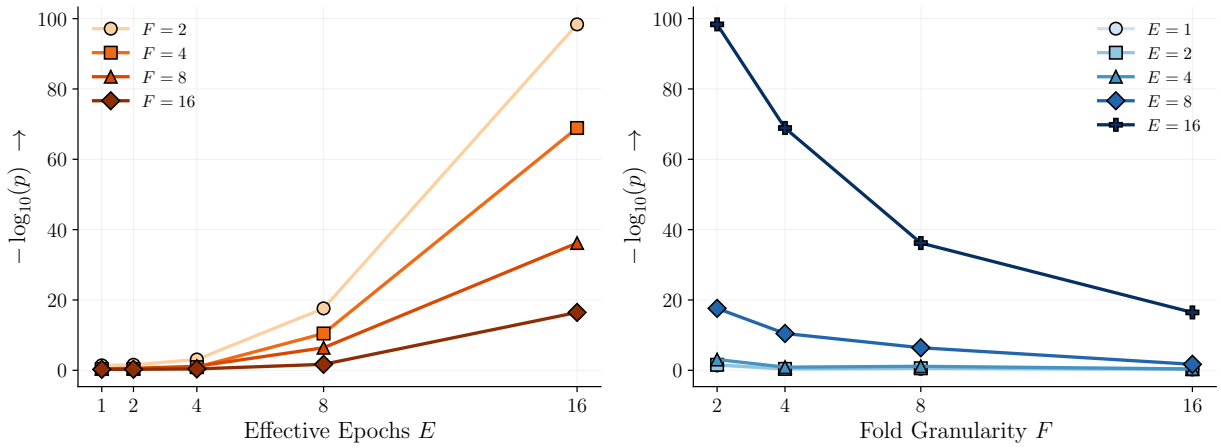


*Figure 20.* Finetuning SKS keyed-signal exposure response across the grid, tracing $-\log_{10} p$ as a function of effective per-key exposure with separate lines for each fold count $F$. This figure plots the same per-cell keyed-signal values as the right panel of Figure 14, just re-cast on a continuous exposure axis to make the trend clearer.

*Table 9.* SKS finetuning: realized exposure summary ($\hat{E}$).

| | $E = 1$ | $E = 2$ | $E = 4$ | $E = 8$ | $E = 16$ |
|---|---|---|---|---|---|
| $F = 2$ | 1.1114 | 2.2492 | 4.5600 | 8.6408 | 16.7533 |
| $F = 4$ | 1.1021 | 2.2335 | 4.6437 | 8.8248 | 17.3686 |
| $F = 8$ | 1.2388 | 2.3070 | 4.6175 | 9.5159 | 17.5870 |
| $F = 16$ | 1.4583 | 2.8750 | 5.5625 | 10.6667 | 20.6875 |

### D.4. SKS Training Scale and Trial Geometry

Table 10 reports the per-cell watermarked-token totals (mean, min–max across paired models) and the corresponding fraction of each model's 131M-token training budget. Table 11 reports the per-cell paired-model counts and the resulting $n_+/n_-$ trial counts; SKS holds the per-cell trial budget fixed at $4/4$ across the grid, since each model trains on exactly one watermarked fold and the per-cell positive/negative budget is the same at every $(F, E)$.

### D.5. SKS Watermark-Only DIA Numeric Table

Table 12 carries the underlying numeric AUC values that drive the SKS DIA heatmap pair in Figure 18. We deliberately keep this table watermark-only: the SKS support construction is a clean per-key scaling ablation in which each model trains on exactly one watermarked fold, and the loss-based / reference-model row-level baselines are only meaningful against the realistic event-split regime where multiple keys coexist, so the head-to-head comparison against those baselines is run only there (Section C.6).

### D.6. SKS Null-Validity

Figure 21 confirms that, once pooled across many distinct positive keys, the empirical exact null is close to uniform at the 1M-null scale and respects standard tail-rate thresholds. This holds in both the event-split grid (Figure 13) and in this SKS ablation, supporting the use of the empirical-null permutation test as the headline reference distribution across both regimes.

## E. Pretraining Exhaustive Readout

This appendix carries the full per-init readout of the pretraining schedule sweep that the main body's Figure 3 compresses into a single combined keyed-signal panel. The structure is grouped first by initialization regime (CPT then from-scratch), and within each by aligned-then-packed surface, exact-then-Gaussian $p$-value type, with the watermark whole-model DIA bar pairs trailing each init's keyed/null block. The split row-level MIA and whole-model DIA baseline tables for both initialization regimes are reported below the per-init bar companions, and each whole-model DIA cell is computed over

*Table 10.* SKS finetuning: training scale context. Per-cell watermark token totals are relative to 131,072,000 train tokens per run. The percent columns report watermark-token share of total train tokens.

| Cell | Target $(E/F)$ | WM tokens mean | WM tokens min-max | Mean % | Range % |
|---|---|---|---|---|---|
| (F2,E1) | 0.5000 | 556,168 | 467,058-671,908 | 0.424% | 0.356%-0.513% |
| (F2,E2) | 1.0000 | 1,133,333 | 962,795-1,167,645 | 0.865% | 0.735%-0.891% |
| (F2,E4) | 2.0000 | 2,281,517 | 2,130,440-2,417,230 | 1.741% | 1.625%-1.844% |
| (F2,E8) | 4.0000 | 4,354,087 | 4,236,298-4,449,342 | 3.322% | 3.232%-3.395% |
| (F2,E16) | 8.0000 | 8,424,968 | 8,120,254-8,677,446 | 6.428% | 6.195%-6.620% |
| (F4,E1) | 0.2500 | 270,914 | 241,723-307,275 | 0.207% | 0.184%-0.234% |
| (F4,E2) | 0.5000 | 547,974 | 503,931-610,453 | 0.418% | 0.384%-0.466% |
| (F4,E4) | 1.0000 | 1,141,527 | 1,028,347-1,249,585 | 0.871% | 0.785%-0.953% |
| (F4,E8) | 2.0000 | 2,164,752 | 2,081,276-2,257,447 | 1.652% | 1.588%-1.722% |
| (F4,E16) | 4.0000 | 4,260,368 | 3,974,090-4,404,275 | 3.250% | 3.032%-3.360% |
| (F8,E1) | 0.1250 | 154,150 | 139,298-176,171 | 0.118% | 0.106%-0.134% |
| (F8,E2) | 0.2500 | 286,790 | 229,432-327,760 | 0.219% | 0.175%-0.250% |
| (F8,E4) | 0.5000 | 573,580 | 540,804-602,259 | 0.438% | 0.413%-0.459% |
| (F8,E8) | 1.0000 | 1,183,009 | 1,052,929-1,249,585 | 0.903% | 0.803%-0.953% |
| (F8,E16) | 2.0000 | 2,203,674 | 2,003,433-2,396,745 | 1.681% | 1.528%-1.829% |
| (F16,E1) | 0.0625 | 71,698 | 57,358-90,134 | 0.055% | 0.044%-0.069% |
| (F16,E2) | 0.1250 | 141,346 | 102,425-176,171 | 0.108% | 0.078%-0.134% |
| (F16,E4) | 0.2500 | 273,475 | 229,432-319,566 | 0.209% | 0.175%-0.244% |
| (F16,E8) | 0.5000 | 524,416 | 471,155-577,677 | 0.400% | 0.359%-0.441% |
| (F16,E16) | 1.0000 | 1,017,080 | 934,116-1,097,996 | 0.776% | 0.713%-0.838% |

*Table 11.* SKS finetuning: model and watermark DIA trial geometry. The WM and clean model columns count target models per cell, and the WM DIA trial column counts the positive/negative pooled trials used for the watermark DIA AUC.

| Cell | WM models $n_+$ | Clean models $n_-$ | WM DIA trials $n_+/n_-$ |
|---|---|---|---|
| (F2,E1) | 4 | 4 | 4 / 4 |
| (F2,E2) | 4 | 4 | 4 / 4 |
| (F2,E4) | 4 | 4 | 4 / 4 |
| (F2,E8) | 4 | 4 | 4 / 4 |
| (F2,E16) | 4 | 4 | 4 / 4 |
| (F4,E1) | 4 | 4 | 4 / 4 |
| (F4,E2) | 4 | 4 | 4 / 4 |
| (F4,E4) | 4 | 4 | 4 / 4 |
| (F4,E8) | 4 | 4 | 4 / 4 |
| (F4,E16) | 4 | 4 | 4 / 4 |
| (F8,E1) | 4 | 4 | 4 / 4 |
| (F8,E2) | 4 | 4 | 4 / 4 |
| (F8,E4) | 4 | 4 | 4 / 4 |
| (F8,E8) | 4 | 4 | 4 / 4 |
| (F8,E16) | 4 | 4 | 4 / 4 |
| (F16,E1) | 4 | 4 | 4 / 4 |
| (F16,E2) | 4 | 4 | 4 / 4 |
| (F16,E4) | 4 | 4 | 4 / 4 |
| (F16,E8) | 4 | 4 | 4 / 4 |
| (F16,E16) | 4 | 4 | 4 / 4 |

$2+/2-$ trials per schedule.

*Table 12.* SKS finetuning: watermark whole-model DIA AUC summary.

| Cell | Aligned Exact | Packed Exact |
|---|---|---|
| (F2,E1) | 0.6875 | 1.0000 |
| (F2,E2) | 0.7500 | 1.0000 |
| (F2,E4) | 0.9375 | 1.0000 |
| (F2,E8) | 1.0000 | 1.0000 |
| (F2,E16) | 1.0000 | 1.0000 |
| (F4,E1) | 0.5625 | 1.0000 |
| (F4,E2) | 0.5000 | 1.0000 |
| (F4,E4) | 0.9375 | 1.0000 |
| (F4,E8) | 1.0000 | 1.0000 |
| (F4,E16) | 1.0000 | 1.0000 |
| (F8,E1) | 0.6875 | 0.8125 |
| (F8,E2) | 0.6875 | 1.0000 |
| (F8,E4) | 0.8125 | 1.0000 |
| (F8,E8) | 1.0000 | 1.0000 |
| (F8,E16) | 1.0000 | 1.0000 |
| (F16,E1) | 0.6875 | 0.6875 |
| (F16,E2) | 0.7500 | 1.0000 |
| (F16,E4) | 0.8750 | 1.0000 |
| (F16,E8) | 1.0000 | 1.0000 |
| (F16,E16) | 1.0000 | 1.0000 |

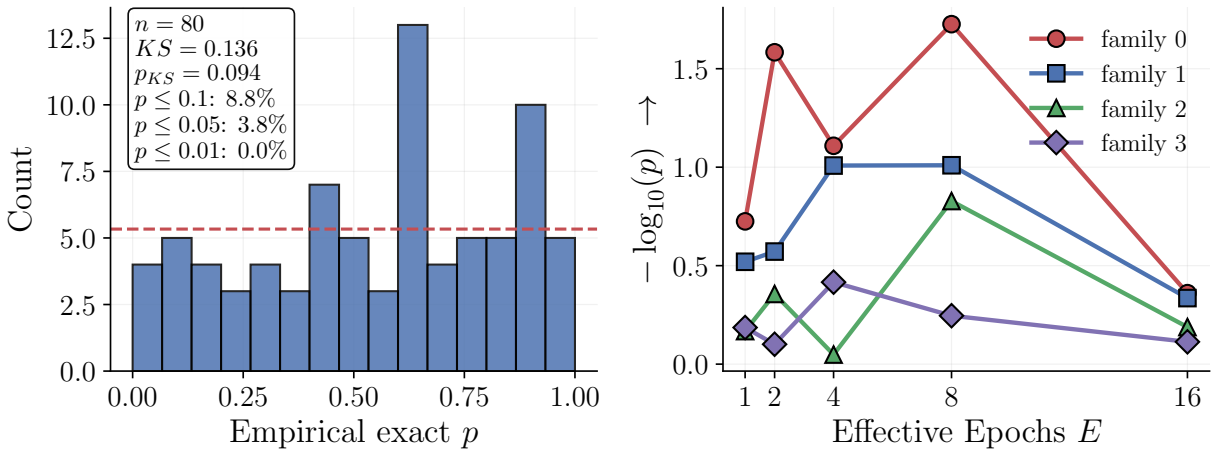


*Figure 21.* SKS null-validity panel. **Left:** histogram of pooled empirical exact $p$-values from the packed clean-model watermark-surface false-probe rows across the depth-4 SKS grid. The annotated KS statistic and $p$-value are from a one-sample Kolmogorov-Smirnov test (`scipy.stats.kstest`) of these pooled $p$-values against $\mathrm{Uniform}(0, 1)$; the dashed horizontal line is the expected per-bin count under a uniform histogram with the plotted binning. **Right:** $-\log_{10} p$ trace of the $F = 2$ warm-key slice of the same packed null family plotted against $E$, shown for context only and not part of the KS test on the left.

*Table 13.* Pretraining idealized exposure profile across the ten-schedule sweep at $F = 2$. Each schedule targets a nominal effective epoch count $E$, which at $F = 2$ corresponds to an idealized per-key exposure $E/F = E/2$. Both initialization regimes (CPT and from-scratch) share the same idealized profile.

| Schedule group | $E$ | $E/F$ |
|---|---|---|
| (S1,E1) – (S4,E1) | 1 | 0.5 |
| (U,E4) / (P,E4) | 4 | 2.0 |
| (U,E8) / (P,E8) | 8 | 4.0 |
| (U,E16) / (P,E16) | 16 | 8.0 |

### E.1. Pretraining CPT Bar Companions

Figures 22 to 25 carry the four CPT keyed/null pairs across the (aligned, packed) surface and (exact, Gaussian) $p$-value-type combinations. Figures 26 and 27 carry the matching CPT watermark whole-model DIA AUC pairs, with aligned and packed surfaces shown side-by-side under each $p$-value type.

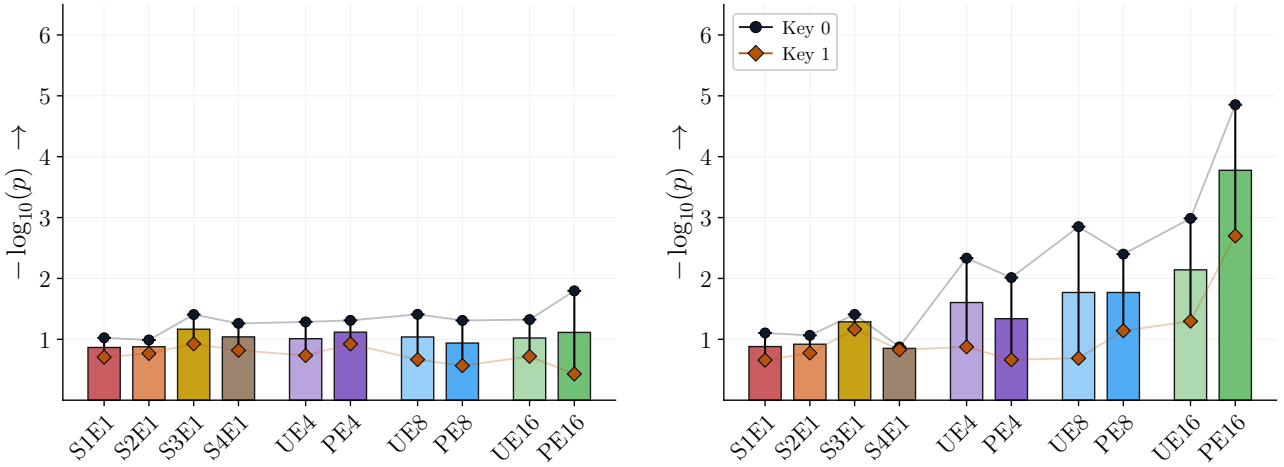


*Figure 22.* Pretraining CPT matched clean-twin false-probe null (left) and keyed signal (right) across the ten-schedule sweep at $F = 2$, on the aligned unpacked detection surface, scored as $-\log_{10} p$ under the exact empirical-null reference. The false-probe null stays in a narrow band across all schedules, while schedule shape clearly modulates the keyed readout, with periodic clusters generally separating more cleanly than the matched uniform variants at $\hat{E} \geq 8$. One of the two inherited keys runs visibly warmer than the other in both panels.

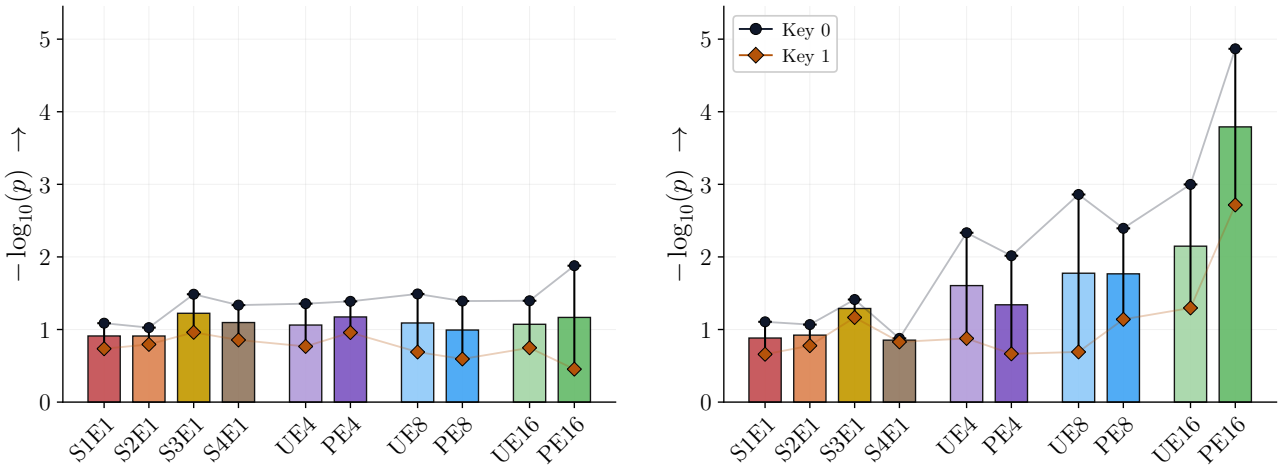


*Figure 23.* Pretraining CPT matched clean-twin false-probe null (left) and keyed signal (right) across the ten-schedule sweep at $F = 2$, on the aligned unpacked detection surface, scored as $-\log_{10} p$ under the empirical-Gaussian reference.

### E.2. Pretraining Scratch Bar Companions

Figures 28 to 31 carry the four from-scratch keyed/null pairs across the (aligned, packed) surface and (exact, Gaussian) $p$-value-type combinations. Figures 32 and 33 carry the matching from-scratch watermark whole-model DIA AUC pairs, with aligned and packed surfaces shown side-by-side under each $p$-value type.

### E.3. Pretraining Schedule Summaries

Tables 14 and 15 provide compact per-schedule summaries combining the realized $\hat{E}$ with the aligned and packed watermark $-\log_{10} p$ values for both initialization regimes.

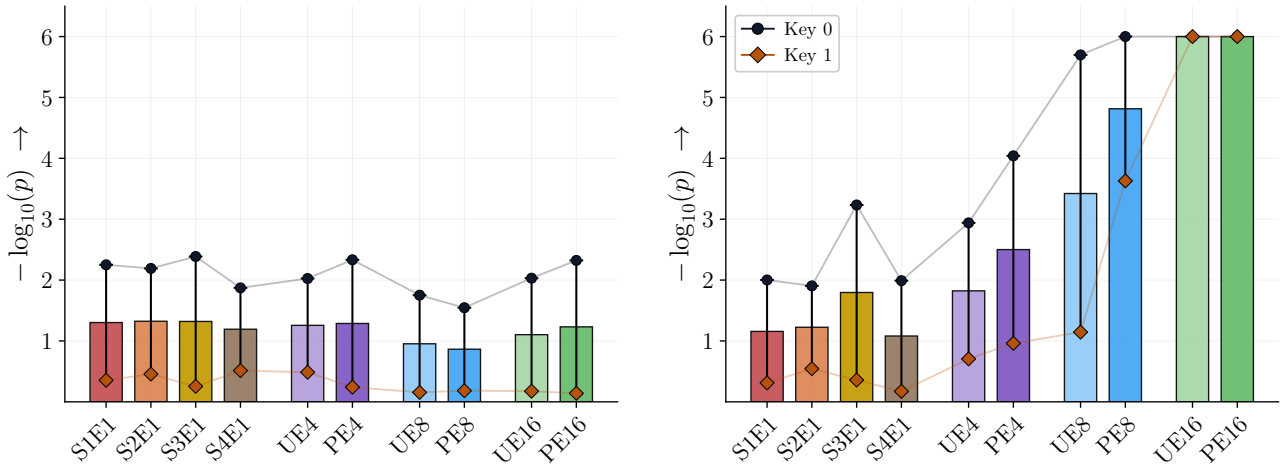


*Figure 24.* Pretraining CPT matched clean-twin false-probe null (left) and keyed signal (right) across the ten-schedule sweep at $F = 2$, on the packed detection surface, scored as $-\log_{10} p$ under the empirical-exact null. The packed surface is a more permissive oracle baseline than the aligned surface.

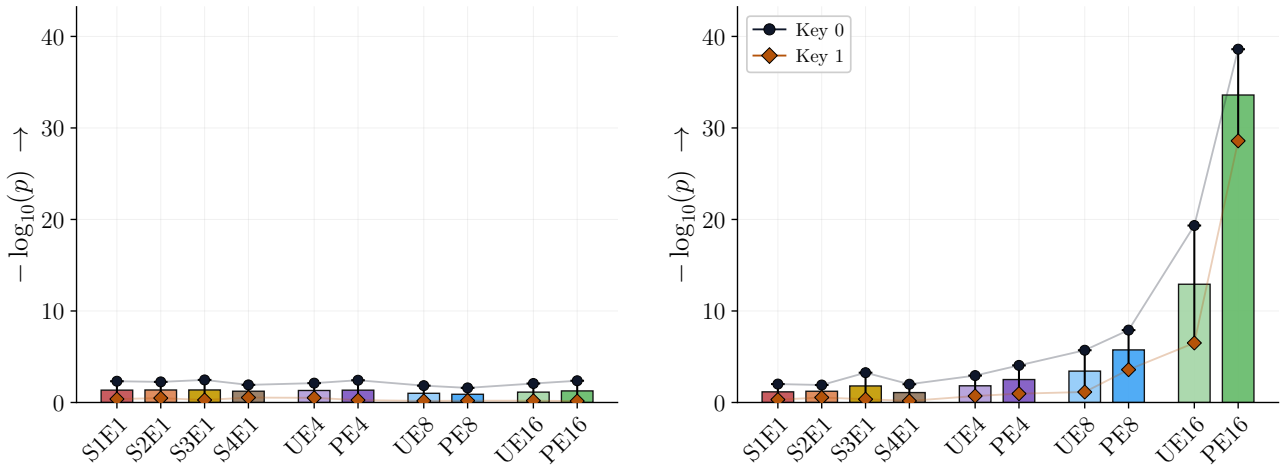


*Figure 25.* Pretraining CPT matched clean-twin false-probe null (left) and keyed signal (right) across the ten-schedule sweep at $F = 2$, on the packed detection surface, scored as $-\log_{10} p$ under the empirical-Gaussian reference.

*Table 14.* CPT pretraining: schedule summary.

| Schedule | Realized ($\hat{E}$) | Aligned WM ($-\log_{10}(p)$) | Packed WM ($-\log_{10}(p)$) |
|---|---|---|---|
| (S1,E1) | 1.0000 | 0.8825 | 1.1630 |
| (S2,E1) | 1.0000 | 0.9225 | 1.2270 |
| (S3,E1) | 1.0000 | 1.2895 | 1.8080 |
| (S4,E1) | 1.0000 | 0.8535 | 1.0840 |
| (U,E4) | 3.8810 | 1.6050 | 1.8240 |
| (P,E4) | 4.0000 | 1.3405 | 2.5080 |
| (U,E8) | 7.8935 | 1.7755 | 3.4280 |
| (P,E8) | 8.0000 | 1.7670 | 5.7460 |
| (U,E16) | 16.2050 | 2.1480 | 12.9210 |
| (P,E16) | 16.0000 | 3.7915 | 33.5980 |

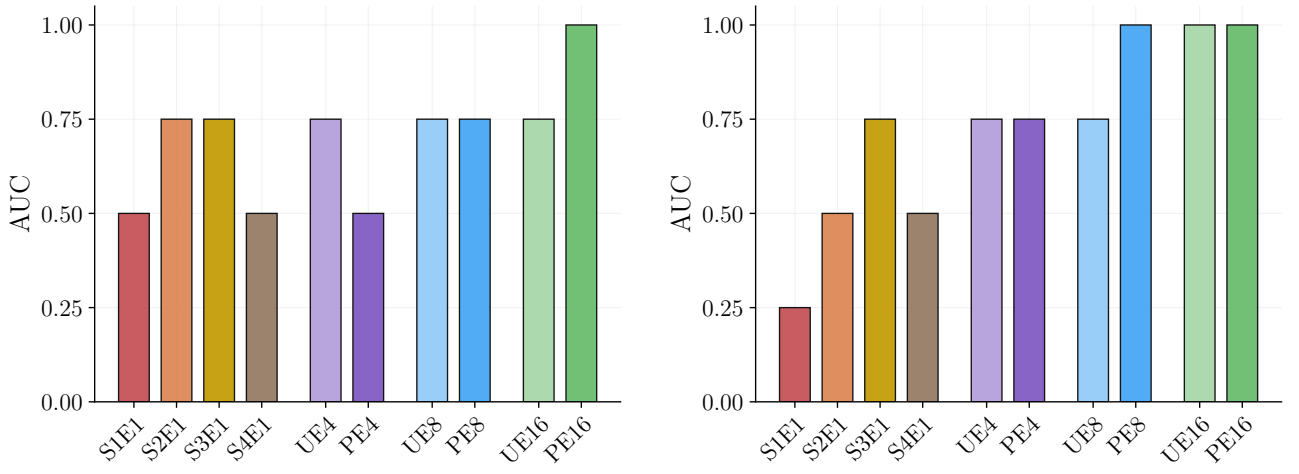


*Figure 26.* Pretraining CPT watermark whole-model DIA AUC across the ten-schedule sweep at $F = 2$, on the aligned (left) and packed (right) detection surfaces, both scored against an empirical-exact null. Each schedule contributes $2+/2-$ whole-model trials, so AUC is coarse but tracks the keyed-signal ordering of Figure 22.

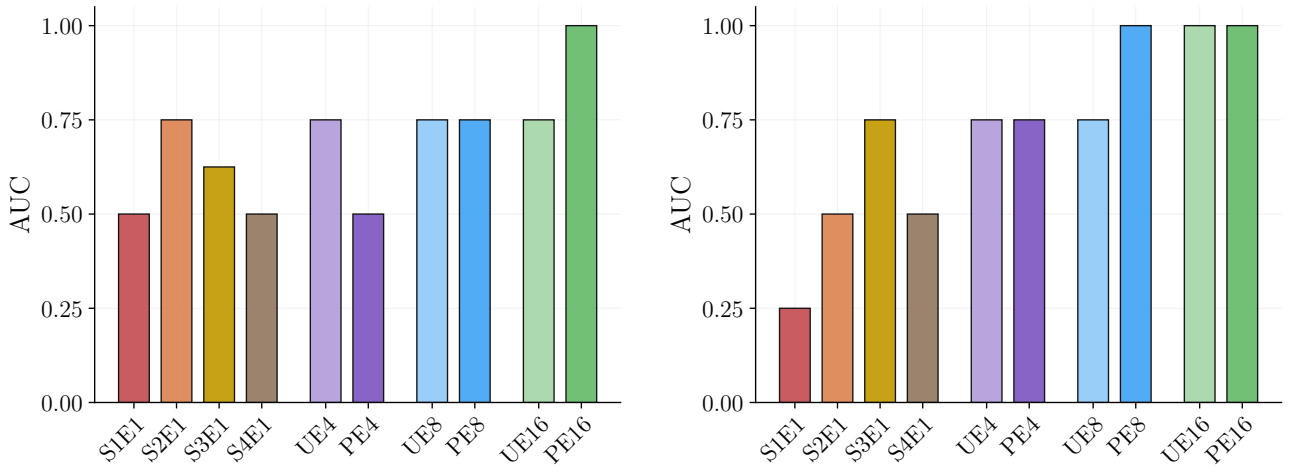


*Figure 27.* Pretraining CPT watermark whole-model DIA AUC across the ten-schedule sweep at $F = 2$, on the aligned (left) and packed (right) detection surfaces, both scored against an empirical-Gaussian null. Each schedule contributes $2+/2-$ whole-model trials.

*Table 15.* From-scratch pretraining: schedule summary.

| Schedule | Realized ($\hat{E}$) | Aligned WM ($-\log_{10}(p)$) | Packed WM ($-\log_{10}(p)$) |
|---|---|---|---|
| (S1,E1) | 1.0000 | 0.4025 | 0.8710 |
| (S2,E1) | 1.0000 | 0.6535 | 1.3460 |
| (S3,E1) | 1.0000 | 0.6415 | 2.7280 |
| (S4,E1) | 1.0000 | 0.8265 | 3.7550 |
| (U,E4) | 3.8895 | 1.5825 | 7.5010 |
| (P,E4) | 4.0000 | 0.9485 | 3.0750 |
| (U,E8) | 7.8115 | 3.9795 | 45.2420 |
| (P,E8) | 8.0000 | 3.2805 | 30.1350 |
| (U,E16) | 16.0740 | 48.4810 | 256.7420 |
| (P,E16) | 16.0000 | 31.4770 | 229.0900 |

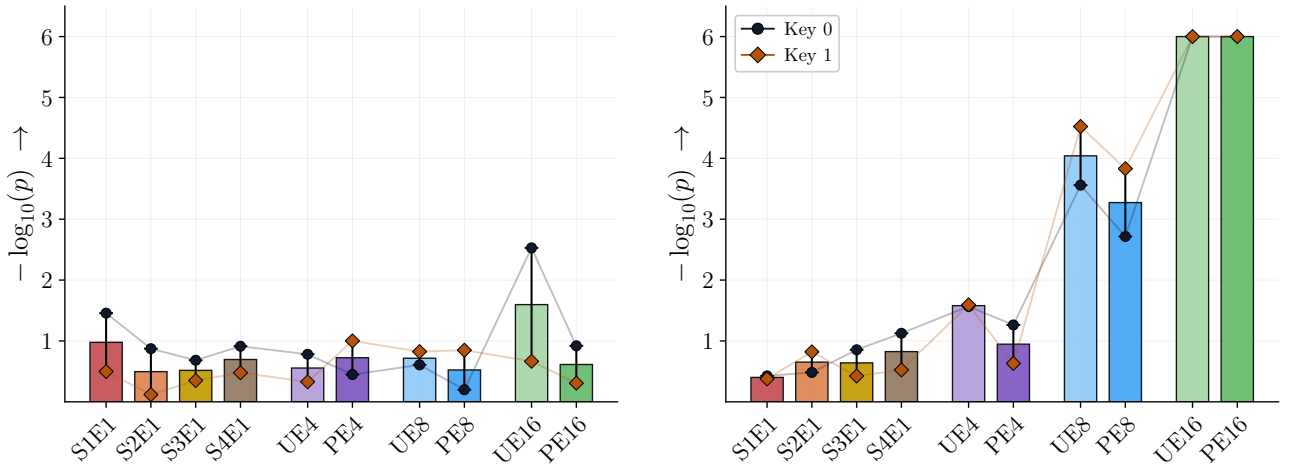


*Figure 28.* Pretraining from-scratch matched clean-twin false-probe null (left) and keyed signal (right) across the ten-schedule sweep at $F = 2$, on the aligned unpacked detection surface, scored as $-\log_{10} p$ under the exact empirical-null reference. The matched clean false-probe null behaves comparably to the CPT case, while the from-scratch keyed readout recovers substantially more strongly at high exposure than the matched CPT runs.

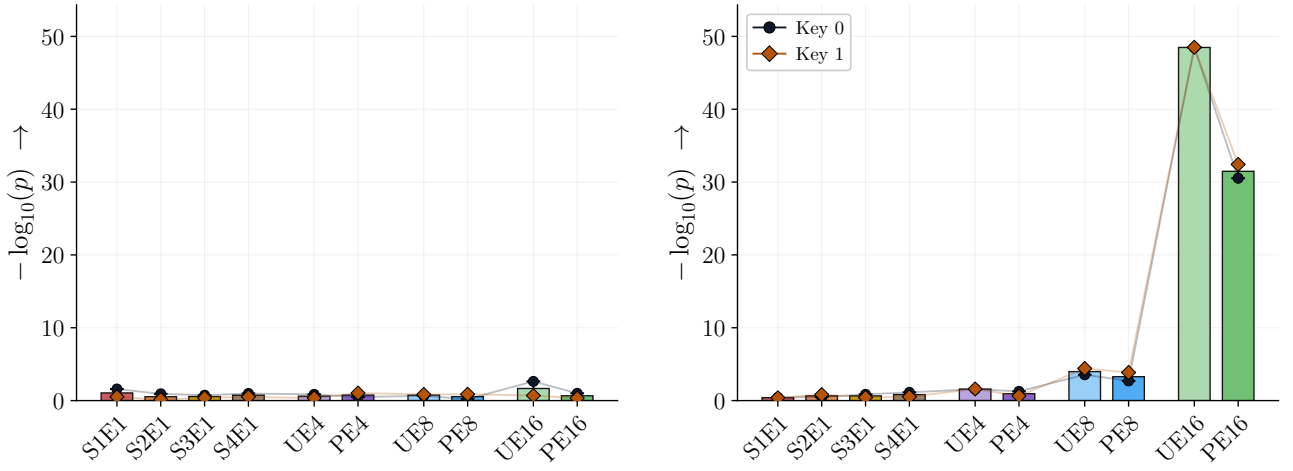


*Figure 29.* Pretraining from-scratch matched clean-twin false-probe null (left) and keyed signal (right) across the ten-schedule sweep at $F = 2$, on the aligned unpacked detection surface, scored as $-\log_{10} p$ under the empirical-Gaussian reference.

### E.4. Pretraining Realized Exposure and $\hat{E}$ Per-Schedule Readbacks

Tables 16 and 17 report the realized normalized exposure $\hat{E}/F$ per schedule for both pretraining initialization regimes, the direct realized counterpart to the idealized $E/F$ values in main-body Table 13. Tables 18 and 19 carry the underlying realized $\hat{E}$ readbacks. The corresponding idealized $E$ targets are the integers in the schedule names (column $E$ of Table 13).

### E.5. Pretraining Training Scale and Trial Geometry

Tables 20 and 21 report the per-schedule watermarked-token totals (mean, min–max across paired models) and the corresponding fraction of each model's 10.49B-token training budget for the CPT and from-scratch initialization regimes respectively. Tables 22 and 23 report the per-schedule paired-model counts and the resulting $n_+/n_-$ trial counts that drive each schedule's whole-model DIA AUC. Each schedule contributes $2/2$ positive/negative whole-model trials per init across both initialization regimes; the in-text pointer in Section 3.2 summarizes the extremal mass values.

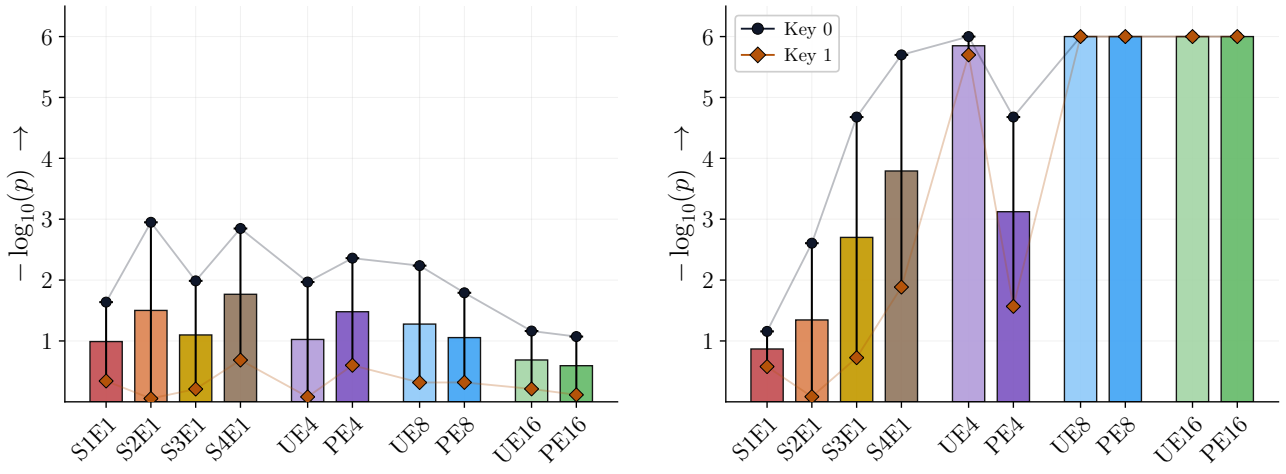


*Figure 30.* Pretraining from-scratch matched clean-twin false-probe null (left) and keyed signal (right) across the ten-schedule sweep at $F = 2$, on the packed detection surface, scored as $-\log_{10} p$ under the empirical-exact null. The packed surface is a more permissive oracle baseline than the aligned surface.

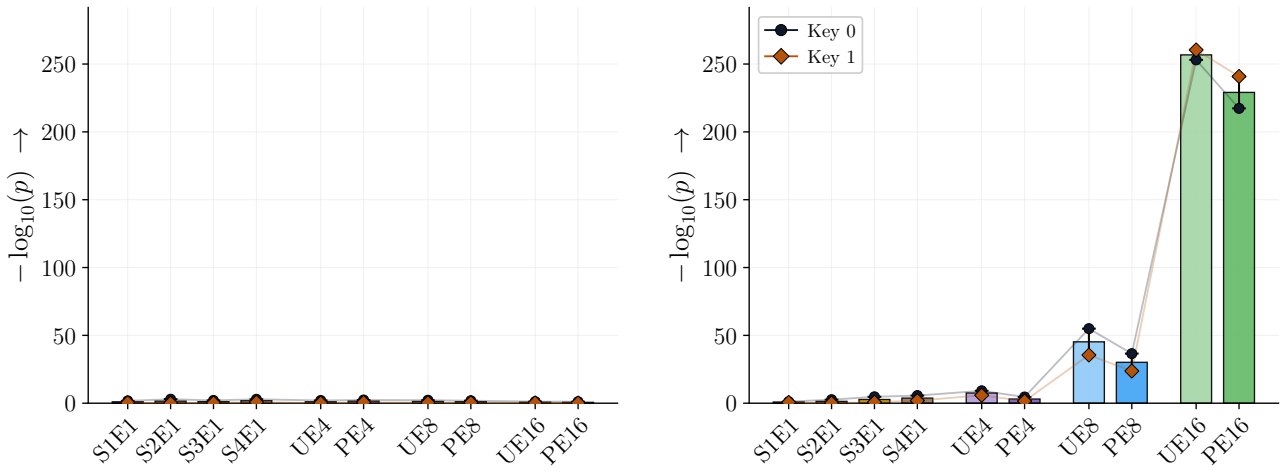


*Figure 31.* Pretraining from-scratch matched clean-twin false-probe null (left) and keyed signal (right) across the ten-schedule sweep at $F = 2$, on the packed detection surface, scored as $-\log_{10} p$ under the empirical-Gaussian reference.

*Table 16.* CPT pretraining: realized normalized exposure summary ($\hat{E}/F$).

| Schedule | Realized ($\hat{E}/F$) |
|---|---|
| (S1,E1) | 0.5000 |
| (S2,E1) | 0.5000 |
| (S3,E1) | 0.5000 |
| (S4,E1) | 0.5000 |
| (U,E4) | 1.9405 |
| (P,E4) | 2.0000 |
| (U,E8) | 3.9467 |
| (P,E8) | 4.0000 |
| (U,E16) | 8.1025 |
| (P,E16) | 8.0000 |

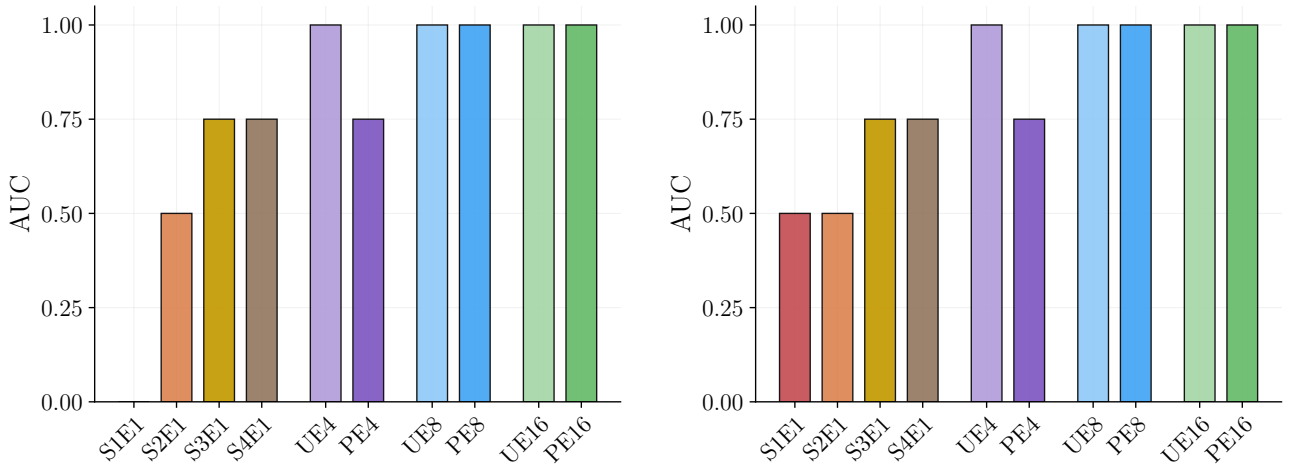


*Figure 32.* Pretraining from-scratch watermark whole-model DIA AUC across the ten-schedule sweep at $F = 2$, on the aligned (left) and packed (right) detection surfaces, both scored against an empirical-exact null. Each schedule contributes $2+/2-$ whole-model trials. The from-scratch DIA recovers near-saturated AUC at the high-exposure schedules well before CPT does.

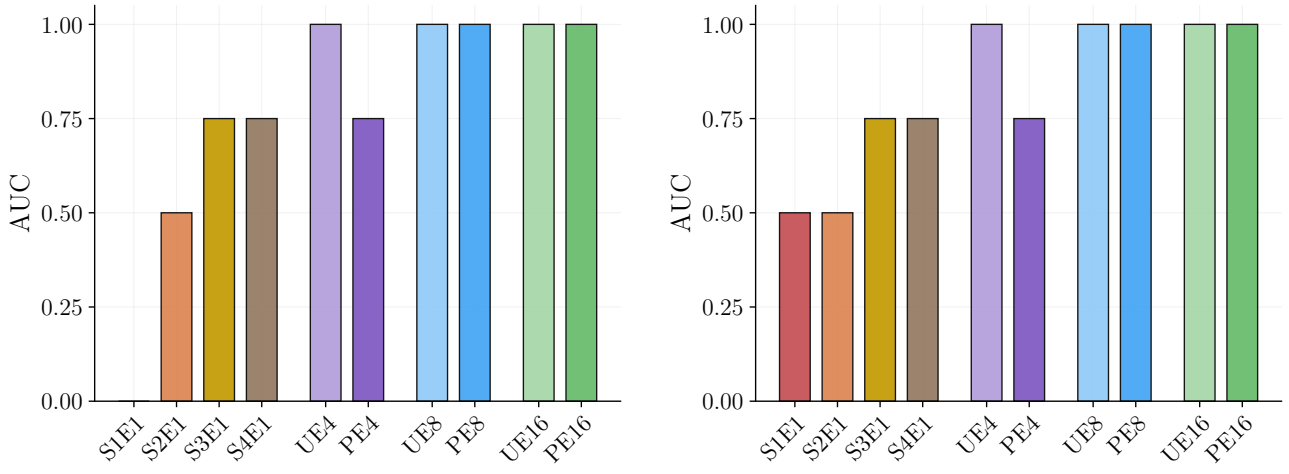


*Figure 33.* Pretraining from-scratch watermark whole-model DIA AUC across the ten-schedule sweep at $F = 2$, on the aligned (left) and packed (right) detection surfaces, both scored against an empirical-Gaussian null. Each schedule contributes $2+/2-$ whole-model trials.

*Table 17.* From-scratch pretraining: realized normalized exposure summary ($\hat{E}/F$).

| Schedule | Realized ($\hat{E}/F$) |
|---|---|
| (S1,E1) | 0.5000 |
| (S2,E1) | 0.5000 |
| (S3,E1) | 0.5000 |
| (S4,E1) | 0.5000 |
| (U,E4) | 1.9447 |
| (P,E4) | 2.0000 |
| (U,E8) | 3.9057 |
| (P,E8) | 4.0000 |
| (U,E16) | 8.0370 |
| (P,E16) | 8.0000 |

*Table 18.* CPT pretraining: realized exposure summary ($\hat{E}$).

| Schedule | Realized ($\hat{E}$) |
|---|---|
| (S1,E1) | 1.0000 |
| (S2,E1) | 1.0000 |
| (S3,E1) | 1.0000 |
| (S4,E1) | 1.0000 |
| (U,E4) | 3.8810 |
| (P,E4) | 4.0000 |
| (U,E8) | 7.8935 |
| (P,E8) | 8.0000 |
| (U,E16) | 16.2050 |
| (P,E16) | 16.0000 |

*Table 19.* From-scratch pretraining: realized exposure summary ($\hat{E}$).

| Schedule | Realized ($\hat{E}$) |
|---|---|
| (S1,E1) | 1.0000 |
| (S2,E1) | 1.0000 |
| (S3,E1) | 1.0000 |
| (S4,E1) | 1.0000 |
| (U,E4) | 3.8895 |
| (P,E4) | 4.0000 |
| (U,E8) | 7.8115 |
| (P,E8) | 8.0000 |
| (U,E16) | 16.0740 |
| (P,E16) | 16.0000 |

*Table 20.* CPT pretraining: training scale context. Per-schedule watermark token totals are relative to 10,485,760,000 train tokens per run. The percent columns report watermark-token share of total train tokens.

| Schedule | WM tokens seen | Realized $\hat{E}$ | Mean % | Range % |
|---|---|---|---|---|
| (S1,E1) | 0.50M | 1.000 | 0.005% | 0.005%-0.005% |
| (S2,E1) | 0.50M | 1.000 | 0.005% | 0.005%-0.005% |
| (S3,E1) | 0.50M | 1.000 | 0.005% | 0.005%-0.005% |
| (S4,E1) | 0.50M | 1.000 | 0.005% | 0.005%-0.005% |
| (U,E4) | 1.94M-1.94M | 3.877-3.885 | 0.019% | 0.019%-0.019% |
| (P,E4) | 2.00M | 4.000 | 0.019% | 0.019%-0.019% |
| (U,E8) | 3.88M-4.01M | 7.762-8.025 | 0.038% | 0.037%-0.038% |
| (P,E8) | 4.00M | 8.000 | 0.038% | 0.038%-0.038% |
| (U,E16) | 8.10M-8.10M | 16.197-16.213 | 0.077% | 0.077%-0.077% |
| (P,E16) | 8.00M | 16.000 | 0.076% | 0.076%-0.076% |

### E.6. Pretraining Loss-Based and Reference-Model Row-Level MIA and Whole-Model DIA Baselines

Tables 24 to 27 report the source-fold loss-based and reference-model row-level MIA and whole-model DIA AUCs alongside the watermark detector's own AUCs (which are the same values that drive Figures 26 and 32).

*Table 21.* From-scratch pretraining: training scale context. Per-schedule watermark token totals are relative to 10,485,760,000 train tokens per run. The percent columns report watermark-token share of total train tokens.

| Schedule | WM tokens seen | Realized $\hat{E}$ | Mean % | Range % |
|---|---|---|---|---|
| (S1,E1) | 0.500M | 1.000 | 0.005% | 0.005%-0.005% |
| (S2,E1) | 0.500M | 1.000 | 0.005% | 0.005%-0.005% |
| (S3,E1) | 0.500M | 1.000 | 0.005% | 0.005%-0.005% |
| (S4,E1) | 0.500M | 1.000 | 0.005% | 0.005%-0.005% |
| (U,E4) | 1.938-1.950M | 3.877-3.902 | 0.019% | 0.018%-0.019% |
| (P,E4) | 1.999M | 4.000 | 0.019% | 0.019%-0.019% |
| (U,E8) | 3.827-3.982M | 7.656-7.967 | 0.037% | 0.036%-0.038% |
| (P,E8) | 3.999M | 8.000 | 0.038% | 0.038%-0.038% |
| (U,E16) | 7.973-8.096M | 15.951-16.197 | 0.077% | 0.076%-0.077% |
| (P,E16) | 7.997M | 16.000 | 0.076% | 0.076%-0.076% |

*Table 22.* CPT pretraining: model and watermark DIA trial geometry. In these pretraining cells, model counts and watermark DIA trial counts coincide because each target contributes one positive or negative whole-model trial to the pooled AUC.

| Schedule | WM models $n_+$ | Clean models $n_-$ | WM DIA trials $n_+/n_-$ |
|---|---|---|---|
| (S1,E1) | 2 | 2 | 2 / 2 |
| (S2,E1) | 2 | 2 | 2 / 2 |
| (S3,E1) | 2 | 2 | 2 / 2 |
| (S4,E1) | 2 | 2 | 2 / 2 |
| (U,E4) | 2 | 2 | 2 / 2 |
| (P,E4) | 2 | 2 | 2 / 2 |
| (U,E8) | 2 | 2 | 2 / 2 |
| (P,E8) | 2 | 2 | 2 / 2 |
| (U,E16) | 2 | 2 | 2 / 2 |
| (P,E16) | 2 | 2 | 2 / 2 |

*Table 23.* From-scratch pretraining: model and watermark DIA trial geometry. In these pretraining cells, model counts and watermark DIA trial counts coincide because each target contributes one positive or negative whole-model trial to the pooled AUC.

| Schedule | WM models $n_+$ | Clean models $n_-$ | WM DIA trials $n_+/n_-$ |
|---|---|---|---|
| (S1,E1) | 2 | 2 | 2 / 2 |
| (S2,E1) | 2 | 2 | 2 / 2 |
| (S3,E1) | 2 | 2 | 2 / 2 |
| (S4,E1) | 2 | 2 | 2 / 2 |
| (U,E4) | 2 | 2 | 2 / 2 |
| (P,E4) | 2 | 2 | 2 / 2 |
| (U,E8) | 2 | 2 | 2 / 2 |
| (P,E8) | 2 | 2 | 2 / 2 |
| (U,E16) | 2 | 2 | 2 / 2 |
| (P,E16) | 2 | 2 | 2 / 2 |

*Table 24.* CPT pretraining: row-level MIA AUC comparison.

| | Watermark Readout | Loss-based Row MIA | | | |
|---|---|---|---|---|---|
| Schedule | WM $-\log_{10}(p_{\text{exact}})$ | Raw-loss | Argmax | min-k$_{10}$ | zlib |
| (S1,E1) | 0.8810 | 0.5029 | 0.5021 | 0.5036 | 0.5023 |
| (S2,E1) | 0.9195 | 0.5165 | 0.5142 | 0.5206 | 0.5127 |
| (S3,E1) | 1.2870 | 0.5291 | 0.5247 | 0.5367 | 0.5225 |
| (S4,E1) | 0.8515 | 0.5079 | 0.5059 | 0.5104 | 0.5062 |
| (U,E4) | 1.6045 | 0.5497 | 0.5457 | 0.5612 | 0.5387 |
| (P,E4) | 1.3390 | 0.5765 | 0.5700 | 0.5939 | 0.5601 |
| (U,E8) | 1.7695 | 0.6104 | 0.6015 | 0.6333 | 0.5893 |
| (P,E8) | 1.7695 | 0.6610 | 0.6521 | 0.6932 | 0.6328 |
| (U,E16) | 2.1415 | 0.7197 | 0.7112 | 0.7540 | 0.6906 |
| (P,E16) | 3.7765 | 0.8163 | 0.8117 | 0.8636 | 0.7884 |

*Table 25.* CPT pretraining: fold-level whole-model DIA AUC comparison. Each cell contributes $2+/2-$ whole-model trials.

| | Watermark DIA | | Loss-based DIA | | | |
|---|---|---|---|---|---|---|
| Schedule | Aligned | Packed | Raw-loss | Argmax | min-k$_{10}$ | zlib |
| (S1,E1) | 0.5000 | 0.2500 | 0.7500 | 0.7500 | 0.7500 | 0.7500 |
| (S2,E1) | 0.7500 | 0.5000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (S3,E1) | 0.7500 | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (S4,E1) | 0.5000 | 0.5000 | 1.0000 | 0.7500 | 1.0000 | 0.7500 |
| (U,E4) | 0.7500 | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (P,E4) | 0.5000 | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (U,E8) | 0.7500 | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (P,E8) | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (U,E16) | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (P,E16) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

*Table 26.* From-scratch pretraining: row-level MIA AUC comparison.

| | Watermark Readout | Loss-based Row MIA | | | |
|---|---|---|---|---|---|
| Schedule | WM $-\log_{10}(p_{\text{exact}})$ | Raw-loss | Argmax | min-k$_{10}$ | zlib |
| (S1,E1) | 0.4015 | 0.5028 | 0.5034 | 0.5064 | 0.5019 |
| (S2,E1) | 0.6520 | 0.5204 | 0.5130 | 0.5245 | 0.5141 |
| (S3,E1) | 0.6395 | 0.5823 | 0.5648 | 0.6058 | 0.5628 |
| (S4,E1) | 0.8250 | 0.6025 | 0.5957 | 0.6517 | 0.5797 |
| (U,E4) | 1.5800 | 0.6846 | 0.6621 | 0.7379 | 0.6525 |
| (P,E4) | 0.9470 | 0.6429 | 0.6043 | 0.6794 | 0.6106 |
| (U,E8) | 4.0410 | 0.8418 | 0.8209 | 0.8848 | 0.8122 |
| (P,E8) | 3.2730 | 0.8472 | 0.8072 | 0.8937 | 0.8131 |
| (U,E16) | 6.0000 | 0.9744 | 0.9768 | 0.9602 | 0.9715 |
| (P,E16) | 6.0000 | 0.9792 | 0.9764 | 0.9711 | 0.9765 |

*Table 27.* From-scratch pretraining: fold-level whole-model DIA AUC comparison. Each cell contributes $2+/2-$ whole-model trials.

| | Watermark DIA | | Loss-based DIA | | | |
|---|---|---|---|---|---|---|
| Schedule | Aligned | Packed | Raw-loss | Argmax | min-$k_{10}$ | zlib |
| (S1,E1) | 0.0000 | 0.5000 | 0.7500 | 1.0000 | 1.0000 | 0.5000 |
| (S2,E1) | 0.5000 | 0.5000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (S3,E1) | 0.7500 | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (S4,E1) | 0.7500 | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (U,E4) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (P,E4) | 0.7500 | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (U,E8) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (P,E8) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (U,E16) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| (P,E16) | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |